The Weighted Nadaraya-Watson Estimator: Strong Consistency Results, Rates of Convergence, and a Local Bootstrap Procedure to Select the Bandwidth

Dissertation
submitted to the Faculty of Economics, Business Administration and Information Technology of the University of Zurich

to obtain the degree of
Doctor of Philosophy
in Banking and Finance

presented by

Kristoph U. Steikert
from Germany

approved in July 2014 at the request of

Prof. Dr. Felix Kübler
Prof. Dr. Michael Wolf

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, July 16, 2014

Chairman of the Doctoral Board: Prof. Dr. Josef Zweimüller

Contents

Acknowledgement **vii**

Preface **ix**

Summary of research results and contributions xii

References xvi

Chapter 1:
The Weighted Nadaraya-Watson Estimator: Pointwise Strong Consistency and Convergence Rates for Strongly Mixing Processes **1**

1.1 Introduction 2

1.2 The weighted Nadaraya-Watson estimator 3

1.3 Main results 8

1.4 Conclusion 18

Appendices 19

A Proofs 19

B Additional lemmas 44

References 47

Chapter 2:
Uniform Strong Consistency and Rates of Convergence for the Weighted Nadaraya-Watson Estimator for Strongly Mixing Processes **51**

2.1 Introduction 52

2.2 The weighted Nadaraya-Watson estimator 54

2.3 Uniform strong consistency 56

2.4 Conclusion 67

Appendices 68

A Proofs 68
B Additional lemmas 99
References 100

Chapter 3:
A Local Bootstrap Procedure to Select the Bandwidth for the Weighted Nadaraya-Watson Estimator in Case of Weakly Dependent Data 103

3.1 Introduction 104
3.2 The estimators 107
3.3 The benchmark method to select the bandwidth 115
3.4 Bandwidth selection via the local bootstrap procedure 116
3.5 Numerical examples 122
3.6 Conclusion 127
Appendix 129
A Tables and Figures 129
References 136

Curriculum Vitae 141

List of Figures

I An illustration of the selection of bootstrap values for the local bootstrap procedure 130
II Comparison of bootstrap data for two different values of the pilot bandwidth 131
III Plots of typical time series data of the considered data generating processes of the simulation study 132
IV Typical scatterplots of lagged data and the true regression functions of the considered models of the simulation study 133
V The resulting box plots of the mean absolute deviation errors (MADEs) of the simulation study 134
VI The resulting kernel density estimates of the MADEs deviation of the simulation study 135

Acknowledgement

I would like to express my deepest gratitude to Felix Kübler and Michael Wolf, my thesis supervisors, for their patient guidance, advise, and useful critiques of this research work. I am also very grateful to my SFI co-advisor Loriano Mancini for his support at the final stage of my studies.

I am thankful to my former colleagues Matthias Jüttner and Jan Wrampelmeyer for making my doctoral studies special.

Last but not least, I would like to thank my parents Annette and Friedrich Steikert for their love, invaluable support, and for believing in me my entire life. Foremost, I wish to thank Silvia Grätz for her love, support, encouragement as well as for her patience throughout my studies and my pursuits.

Zurich, 2014 Kristoph U. Steikert

Preface

This dissertation consists of three research manuscripts establishing strong consistency results as well as introducing a bandwidth selection method for the weighted Nadaraya-Watson estimator. This nonparametric estimator, introduced in Hall et al. (1999) and subsequently investigated in Cai (2001), extends the ordinary Nadaraya-Watson estimator. It is designed to reproduce the superior bias properties of local linear methods, while, in the case of estimating the predictive distribution, preserving the property that the ordinary Nadaraya-Watson estimator is always a distribution function. Therefore, in the context of weakly dependent, nonlinear data, the estimator is appealing, because, compared to classical methods, it provides a coherent and bias reduced framework for forecasting time series.

Estimating a future state given a selective past (lagged variables) is important but provides only a fragment of the information necessary in order to forecast. The predictive distribution on the other hand contains all information about this future state given the selective past. Correctly estimating this conditional cumulative distribution function (CDF) is therefore essential in a serious attempt to forecast. The weighted Nadaraya-Watson estimator ensures that estimates exhibit the superior bias properties compared to the corresponding ordinary Nadaraya-Watson estimates. The bias properties are the same as the ones of the local linear estimator but this estimator suffers from not being a proper estimator of the conditional CDF, because negative probabilities cannot be ruled out. Also, for estimating the predictive variance, e.g., to provide confidence intervals for estimates of a multi-step-ahead prediction, nonnegative values may occur. The weighted Nadaraya-Watson estimator, however, always produces nonnegative estimates. It therefore covers the entire range of statistics necessary to forecast while maintaining favorable bias properties and provides a coherent and sound approach to forecast nonlinear time series.

For forecasting to be meaningful the underlying point estimators must be point-

wise consistent, i.e., the estimator tends to the quantity being estimated as sample size increases indefinitely. For large samples this translates into a certain amount of confidence that the point estimate is close to the true value, or put differently, that the estimation error is small. This confidence is expressed through the concepts of weak and strong consistency. The notion of weak consistency is that as the sample size increases indefinitely, the probability of the estimator being arbitrarily close to the true value approaches one. Strong consistency, on the other hand, states that the probability of the event "the estimator converges to the true value" is equal to one. That is, all events for which the estimator does not converge to the true value have zero probability. The differences between the two concepts are substantial, because weak consistency is a notion about probabilities of particular events, namely "the estimator is arbitrary close to the value being estimated", whereas the notion of strong consistency demands with probability one that the deviation becomes small and remains small. Thus, it is equivalent to the classical (real analysis) notion of convergence for almost all sequences of the estimator implying that there exists an (unknown) size of the sample for which the estimator is arbitrary close to the true value. If the estimator is weak but not strong consistent, then it may occur that for a larger sample a larger estimation error is realized than for a smaller sample. To prove weak consistency it suffices, by virtue of Chebyshev's inequality, to establish that the estimator is asymptotically unbiased and its variance approaching zero. In contrast, the workhorse for establishing strong consistency is the Borel-Cantelli lemma (see Resnick (2003, p. 102)). The conditions of the lemma involve the summability of the probabilities of the event "the estimator deviates from the value being estimated by an arbitrary threshold" as sample size increases indefinitely. To prove the finiteness of this infinite sum of probabilities requires (usually) more effort than proving the above two requirements for weak consistency in a comparable setting.

Pointwise consistency is an important property, however, to extend the analysis and to be able to prove further consistency results for estimators in which the weighted Nadaraya-Watson estimator is embedded, it does not suffice. This is because this notion of consistency does not rule out the possibility that consistency breaks down at certain points of the domain close to the point at which pointwise consistency holds. To resolve this deficit, a stronger notion of consistency, namely uniform consistency, is needed. An estimator is uniform consistent on a compact subset of $\mathbb{R}^d$, for $d \geq 1$, if the sequence of uniform (or supremum) norms of the difference between the estimator and the value being estimated, defined on the compact subset of the domain, converges to zero. In other words, the maximal estimation

error on this subset convergence to zero. Similar as in the pointwise case notions of weak and strong uniform consistency apply depending on the notions of convergence of random variables.

In this dissertation a canonical approach for establishing strong consistency for the weighted Nadaraya-Watson estimator is followed. Since the estimator is a point estimator, i.e., given the observed data the computed statistic is single valued, pointwise (strong) consistency is a natural first property that needs to be established. The first manuscript of this dissertation therefore establishes pointwise strong consistency for the weighted Nadaraya-Watson estimator and provides rates of convergence. These rates measure the speed for which the sequence of estimators approaches the true value. They depend on the bandwidth and the sample size. The results of the manuscript are novel because pointwise strong consistency for this estimator has not been established before. A detailed summary of the manuscript is given in the next section.

For a deeper understanding and further applications, however, pointwise consistency does not suffice and therefore establishing strong uniform consistency is the next consequential step in the analysis of consistency properties. The second manuscript of this dissertation therefore establishes uniform strong consistency of the weighted Nadaraya-Watson estimator, on compact subsets of the real line, and provides rates of convergence. The result is novel because it constitutes the first result regarding uniform consistency for this estimator. The manuscript provides the foundations necessary to prove further consistency results of estimators in which the weighted Nadaraya-Watson estimator is embedded, such as two-step, semiparametric, and bootstrap estimators which are applied in various fields. The result thereby completes the fundamental analysis about consistency of this estimator. A detailed summary of the manuscript is given in the next section.

The weighted Nadaraya-Watson estimator is a local estimator in the sense that data in the neighborhood of the point of evaluation in the predictor space are combined to produce an estimate for this point. For each estimate it is therefore essential to determine the size of this neighborhood which is represented by the bandwidth. This free parameter exhibits a strong influence on the resulting estimate. In particular, a large bandwidth is associated with a larger bias whereas a small bandwidth increases the variance. Thus, a fundamental task in determining the bandwidth is to balance this trade-off to filter the signal from the noise. Selecting the bandwidth is therefore a crucial task in the estimation process. However, methods to select the bandwidth for the weighted Nadaraya-Watson estimator, in particular

in the case of weakly dependent time series data, are scarce. The third and final manuscript proposes a bootstrap estimator of the integrated mean squared error (IMSE) of the weighted Nadaraya-Watson estimator to select the bandwidth based on the local bootstrap. Since the MSE separates into squared bias and variance of the estimator the estimated MSE represents a convenient pointwise loss function. The bandwidth is selected such that it minimizes the integrated version of this loss function. The procedure is tested in an extensive simulation study estimating various statistics of future values based on observed values for commonly used nonlinear time series models. In terms of the mean absolute deviation error (MADE) the procedure outperforms a selection procedure based on a nonparametric version of Akaike's information criterion that is frequently used in the applied literature. The results show that the proposed procedure based on the local bootstrap is an appealing choice among the scarce list of bandwidth selection methods for the weighted Nadaraya-Watson estimator and weakly dependent time series data.

Summary of research results and contributions

This dissertation consists of the following three research manuscripts:

- The Weighted Nadaraya-Watson Estimator: Pointwise Strong Consistency and Convergence Rates for Strongly Mixing Processes
- Uniform Strong Consistency and Rates of Convergence for the Weighted Nadaraya-Watson Estimator for Strongly Mixing Processes
- A Local Bootstrap Procedure to Select the Bandwidth for the Weighted Nadaraya-Watson Estimator in Case of Weakly Dependent Data

Their content and contribution are summarized in the following subsections.

The Weighted Nadaraya-Watson Estimator: Pointwise Strong Consistency and Convergence Rates for Strongly Mixing Processes

This manuscript establishes pointwise strong consistency for the weighted Nadaraya-Watson estimator for functions of strongly mixing processes. Considering functions of strongly mixing processes facilitate the estimation of the entire range of statistics needed to forecast including the multi-step ahead prediction given a selective

past, raw moments and the variance thereof, as well as its distribution function (predictive distribution). The consistency result established in this manuscript is stronger than the weak consistency result provided by Cai (2001). This work therefore completes the classical analysis about pointwise consistency for this particular estimator. Furthermore, this manuscript provides the convergence rate which is novel for the weighted Nadaraya-Watson estimator. This rates, which measures the speed at which the convergent sequence of estimators approaches its limit, coincides with rates for the ordinary Nadaraya-Watson estimator previously established in Cheng (1995, pp. 361–362) and Sarda and Vieu (2000, p. 62). While in the aforementioned papers the underlying processes are assumed to be independent, the same convergence rate is achieved in this manuscript although the weighted Nadaraya-Watson estimator is a constrained kernel estimator and the underlying processes are assumed to be weakly dependent.

Uniform Strong Consistency and Rates of Convergence for the Weighted Nadaraya-Watson Estimator for Strongly Mixing Processes

To extend the consistency analysis of the weighted Nadaraya-Watson estimator this manuscript establishes uniform strong consistency over compact subsets of the real line. Similar to the previous manuscript functions of strongly mixing processes are considered. Uniform strong consistency is an important property of estimators, because it permits further research regarding consistency of estimation methods in which the weighted Nadaraya-Watson estimator is embedded. Examples include two-stage, semiparametric, or bootstrap estimators. The results in this manuscript therefore provide the foundations for proving consistency of the bootstrap estimator in the final manuscript of this dissertation.

To emphasize the importance of uniform consistency consider the following simple example. Let $T_n = T_n(X_1, X_2, ..., X_n) = \sqrt{nh}(\widehat{m}(x) - m(x))$ where $\widehat{m}(x)$ is the weighted Nadaraya-Watson estimator of $m(x)$. The sampling distribution of T_n is unknown because the underlying data are a random sequence with unknown joint CDF, denoted by $F_0 \in \mathcal{F}$. Furthermore, let $G_n(\cdot, F)$ denote the exact sample distribution of T_n when the underlying data are sampled from F, where F is a generic element of the class, $\mathcal{F}$, of finite-dimensional and continuous CDFs. To estimate the sampling distribution $G_n(\cdot, F_0)$ the bootstrap approach, introduced by Efron (1979), replaces the unknown exact CDF, F_0, by a consistent estimator $\widehat{F}_n$. The bootstrap estimator $G_n(\cdot, \widehat{F}_n)$ is consistent if it is uniformly close to the asymptotic

CDF of T_n, denoted by $G_\infty(\cdot, F_0)$, for large n. Thus, for the bootstrap estimator to be consistent $G_n(\cdot, \widehat{F}_n)$ must converge uniformly to $G_\infty(\cdot, F_0)$.

The result presented in the manuscript is the first uniform consistency result for the weighted Nadaraya-Watson estimator. Due to the strong nature of the result it implies weak uniform consistency over compact subsets of the real line. Furthermore, this manuscript provides rates of convergences. This rate is optimal in the sense of Stone (1982, Theorem 1). For practical purposes a detailed analysis of the rate of convergence in the case of a polynomial strong mixing coefficient is added.

A Local Bootstrap Procedure to Select the Bandwidth for the Weighted Nadaraya-Watson Estimator in Case of Weakly Dependent Data

The selection of the bandwidth for nonparametric estimators such as the weighted Nadaraya-Watson estimator and the class of local polynomial estimators is the most crucial task of the estimation process. The bandwidth, being a free parameter, strongly influences the resulting estimate. The bias and variance of the estimator are positively respectively negatively related to the bandwidth. The fundamental task of the selection process is therefore to choose a bandwidth that balances the bias and variance of the estimator in order to filter the signal from the noise.

Methods to select the bandwidth for weakly dependent time series data are scarce. Existing methods based on the autoregressive bootstrap postulate enough prior knowledge about the data generating process to select the bandwidth accurately. A selection according to a nonparametric version of Akaike's information criterion, introduced by Cai and Tiwari (2000), is easier to implement than the aforementioned procedure but, as this study shows, systematically produces unsatisfactory results.

This manuscript proposes a novel, fully data driven method to select the bandwidth for the weighted Nadaraya-Watson estimator. The procedure is based on the local bootstrap proposed by Paparoditis and Politis (2000). They introduce the procedure to approximate the sampling distribution of kernel estimators, in particular they consider the Nadaraya-Watson estimator. For the present work their approach is extended to the weighted Nadaraya-Watson estimator due to its favorable bias properties. The procedure extends because the estimator consistently estimates the conditional CDF of a future event given a selective past. Given the local bootstrap an estimator of the integrated mean squared error (IMSE) is constructed and a particular bandwidth is selected for which the estimated IMSE is minimized. In

general the IMSE is a convenient function to determine the bandwidth because the mean squared error is equivalent to the sum of squared bias and the variance of the weighted Nadaraya-Watson estimator. Due to the consideration of the local bootstrap an additional bandwidth, the so called pilot or resampling bandwidth, emerges. The selection of bandwidth depends on the choice of this pilot bandwidth. The dependence issue is dampened by introducing an iterated bandwidth selection scheme and simple choices for the initial pilot bandwidth.

The advantages of the bootstrap procedure are twofold. First, it is a local procedure in the sense that given the selective past only future observations are bootstrapped. This implies that the procedure is easy to implement and is capable of coping with a large sample size as well as with a large number of bootstrap samples. Second, the procedure is entirely nonparametric, i.e., it avoids any assumption regarding the parametric form of the data generating process. This is appealing because unlike in the case of the autoregressive bootstrap no parametric form for the underlying nonlinear data generating process is specified.

The selection procedure is tested in an extensive simulation study estimating various statistics of future values based on observed values for commonly used nonlinear time series models. The performance of the selection procedure is measured by the mean absolute deviation error (MADE) which is frequently used in the literature. For given bandwidth it measures the mean absolute difference between the estimates and the value being estimated given a set of selected pasts. The results show that the procedure outperforms a benchmark procedure based on a nonparametric version of Akaike's information criterion that is frequently used in the literature. Furthermore, the MADEs, given the bandwidth selected via the local bootstrap procedure, are close to the MADEs given the empirically optimal bandwidth. This bandwidth is optimal in a squared sense, minimizing the squared difference of the estimates and the values being estimated given a set of selected pasts. The results show that the proposed selection procedure is an appealing choice among the scarce list of bandwidth selection methods for the weighted Nadaraya-Watson estimator in case of weakly dependent time series data.

References

Cai, Z. (2001): "Weighted Nadaraya-Watson Regression Estimation," *Statistics & Probability Letters*, 51, 307–318.

Cai, Z. and R. C. Tiwari (2000): "Application of a Local Linear Autoregressive Model to BOD Time Series," *Environmetrics*, 11, 341–350.

Cheng, P. E. (1995): "A Note on Strong Convergence Rates in Nonparametric Regression," *Statistics & Probability Letters*, 24, 357–364.

Efron, B. (1979): "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

Hall, P., R. C. L. Wolff, and Q. Yao (1999): "Methods for Estimating a Conditional Distribution Function," *Journal of the American Statistical Association*, 94, 154–163.

Paparoditis, E. and D. N. Politis (2000): "The Local Bootstrap for Kernel Estimators under General Dependence Conditions," *Annals of the Institute of Statistical Mathematics*, 52, 139–159.

Resnick, S. I. (2003): *A Probability Path*, Boston: Birkhäuser, 3rd ed.

Sarda, P. and P. Vieu (2000): "Kernel Regression," in *Smoothing and Regression: Approaches, Computation, and Application*, ed. by M. G. Schimek, New York: John Wiley & Sons, chap. 3, 43–70.

Stone, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053.

Chapter 1

The Weighted Nadaraya-Watson Estimator: Pointwise Strong Consistency and Convergence Rates for Strongly Mixing Processes

1.1 Introduction

For point estimators such as the weighted Nadaraya-Watson estimator pointwise consistency (henceforth consistency) is probably the most important property. An estimator is consistent if the estimate tends to the true value (the value being estimated) as the size of the sample increases indefinitely. Thus, for large samples a consistent estimator provides a certain amount of confidence that the estimate is close to the true value.

A consistent estimator is either weak consistent, strong consistent, or both. Heuristically the notion of weak consistency is that the sequence of distributions of the estimates become increasingly concentrated around the true value, so that the probability of the estimator being arbitrarily close to true value converges to one. Because of its probabilistic nature there exists even for large samples positive probability that the estimate deviates from the quantity being estimated. Strong consistency on the other hand states that the probability of the event "the sequence of estimators converges to the true value" is one. It is equivalent to the classical (real analysis) notion of convergence for almost all sequences of the estimator and is therefore a stronger concept than the latter.

In this manuscript I establish strong consistency for the weighted Nadaraya-Watson estimator for weakly dependent processes under fairly weak conditions. The result is stronger that the pointwise weak consistency result provided by Cai (2001). Furthermore, I provide the rate for which the estimator converges, i.e., the speed at which the sequence of estimators converges almost surely to the true value.

The weighted Nadaraya-Watson estimator, introduced in Hall and Presnell (1999) and generalized in Cai (2001), combines favorable features of the Nadaraya-Watson and local linear estimator, respectively. First, it is a proper estimator of the conditional cumulative distribution function (CDF) which the local linear is not (see Yu and Jones (1998) and Hall et al. (1999)). Second, it exhibits the same finite-sample bias as the local linear estimator. Compared to the Nadaraya-Watson estimator this bias is favorable. The results in this manuscript extend the literature of asymptotic properties of nonparametric estimators. In particular, for the weighted Nadaraya-Watson estimator in the case of weakly dependent data and therefore completes the analysis of pointwise consistency.

Strong consistency for the ordinary Nadaraya-Watson estimator has been established in, e.g., Cheng (1995), Sarda and Vieu (2000), and Walk (2010). The convergence rate in this manuscript coincide with the rates previously established

in Cheng (1995, pp. 361–362) and Sarda and Vieu (2000, p. 62). This is interesting because the aforementioned studies assume independent data whereas I consider weakly dependent data. In addition, although the weighted Nadaraya-Watson is a constrained Nadaraya-Watson estimator, fulfilling certain bias properties, it achieves the same rate of convergence as the ordinary Nadaraya-Watson estimator. Walk (2010) establishes strong consistency using a weaker set of assumptions but does not provide the rate of convergence.

Regarding data dependence I assume an univariate time series framework where the considered processes are strongly mixing. Strong mixing, introduced in Rosenblatt (1956), is a widely adopted assumption in the literature. As all mixing concepts strong mixing indicates the maximum dependence between two time time events at least some steps apart (see Doukhan (1994) for an in-depth treatment). If this dependence converges to zero it indicates a notion of asymptotic independence and the sequence is said to be strongly mixing. Examples of strongly mixing time series are finite-dependent processes, types of ARMA processes (see Davidson (1994, pp. 219–228) for sufficient conditions), classes of Markov chains (Bradley (2005, pp. 117–122)), as well as linear GARCH processes (see Basrak et al. (2002, theorem 3.1) and the reference therein as well as Lindner (2009) for an overview of the probabilistic properties of GARCH) and (non-)stationary ARCH processes (Fryzlewicz and Subba Rao (2011)).

Although a time series setting is considered all the established results hold for a more general setting. I will highlight the differences throughout the text if necessary.

The remainder of the manuscript is organized as follows. The next section introduces the weighted Nadaraya-Watson estimator. Section 1.3 presents the main results, including a detailed discussion of the underlying assumptions to establish pointwise strong consistency of the estimator. All proofs are provided in Appendix A. In addition, Appendix B provides additional lemmas necessary to prove the main result.

1.2 The weighted Nadaraya-Watson estimator

Let $\{X_i\}_{i=1}^{n+1}$ to be a strictly stationary real valued time series defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\phi(\cdot)$ denote an arbitrary Borel-measurable

function. The regression function $m(\cdot)$ at location x is defined as

$$m(x) = \mathbb{E}(\phi(X_{i+1})|X_i = x), \tag{1.1}$$

assuming $\mathbb{E}\,|\phi(X_{i+1})| < \infty$. Thus, $m(x)$ is the conditional expectation of $\phi(X_{i+1})$ given the previous observation X_i realizes to x. This is equivalent to the best mean squared prediction of $\phi(X_{i+1})$ based on the information $X_i = x$. Introducing $\phi(\cdot)$ as the response function provides an enriched flexibility in expressing various statistics such as the simple one-step ahead prediction ($\phi(X_{i+1}) = X_{i+1}$), raw moments thereof ($\phi(X_{i+1}) = X_{i+1}^j$ for integers $j > 0$), or conditional probabilities ($\phi(X_{i+1}) = \mathbb{1}_{(-\infty,y]}(X_{i+1})$ for some $y \in \mathbb{R}$). The conditioning variable can be any lagged variable by replacing $X_i = x$ with $X_{i-j} = x$, where the integer j is finite, nonnegative, and constraint by the number of observations, or a vector thereof. To keep notation parsimonious $m(x)$ is a function with a one-dimensional input as in (1.1) but it needs to be emphasized that the extension to multi-dimensional inputs is straightforward.

Note that I focus on a time series framework. This setting is a special case of a more general setting. All the presented results hold for processes of the form $\{(X_i, Y_i)\}_{i=1}^n$ because by setting $Y_i = X_{i+1}$ the presented framework emerges. In Section 1.3.1 I therefore add the assumptions necessary to prove the results for the general case. In what follows and in particular in the proofs below I complement the text for this case if necessary.

To estimate (1.1), nonparametric estimation, compared to parametric estimation, omits to postulate a specific global structure regarding the regression function $m(\cdot)$. Instead it approximates the local structure of $m(\cdot)$ at the point x, where x is surrounded by or equal to observations in the sample which all provide a certain amount of information to determine the estimate $\widehat{m}(x)$ of $m(x)$. The locality is incorporated by assuming that data in a close neighborhood to x contain more information about the regression function than data farther away. To formalize the idea assume that the $(q+1)$-th derivative of $m(\cdot)$ at x exists. Then approximate $m(X_i) = \mathbb{E}(\phi(X_{i+1})|X_i)$, locally at x, given a prespecified size of the local neighborhood denoted by h_n, by a polynomial of total order q. That is,

$$m(X_i) \approx \sum_{k=0}^{q} \frac{1}{k!} m^{(k)}(x)(X_i - x)^k, \tag{1.2}$$

with $m^{(k)}(x)$ denoting the k-th derivative of $m(\cdot)$ evaluated at x. The model param-

eters, here $m^{(k)}(x)/k!$, are local parameters and therefore depend on x. Estimation of the local model is conducted using a locally weighted polynomial regression, i.e., the estimator emerges by minimizing

$$\sum_{i=1}^{n}\left(\phi(X_{i+1})-\sum_{k=0}^{q}\frac{1}{k!}m^{(k)}(x)(X_i-x)^k\right)^2 p_i(x;\lambda_n)K_{h_n}(X_i-x), \tag{1.3}$$

with respect to $m^{(k)}(x)/k!$, $k=0,1,...,q$. The functions $K_{h_n}(u)=K(u/h_n)/h_n$ and $p_i(x;\lambda_n)$ are both nonnegative weight functions. The kernel function $K:\mathbb{R}\to\mathbb{R}$ controls the degree of information determining the estimate of $m(x)$. Provided $K(\cdot)$ is a unimodal symmetric density, as is the case in the present manuscript, it down-weights contributions of data farther apart from x. The probabilities $p_i(x;\lambda_n)$, which are unique to the weighted Nadaraya-Watson estimator, constrain the estimator to fulfill a linearity condition implying a favorable finite-sample bias which is discussed below in more detail.

It is important to note that the estimation process is more or less typical to a regression analysis by considering the variable in its state rather than in its time domain. This is explained by the whitening by windowing principle (see Hart (1996, pp. 117–119)) which will become evident in the proofs below.

To introduce the ordinary Nadaraya-Watson estimator suppose that the probabilities are uniform, i.e., $p_i(x;\lambda_n)=n^{-1}$ for all i. If the approximation in (1.2) is constant, i.e., $q=0$, then the estimator for $m(x)$ reads

$$\widehat{m}_{nw}(x)=\frac{\sum_{i=1}^{n}K_{h_n}(X_i-x)\phi(X_{i+1})}{\sum_{i=1}^{n}K_{h_n}(X_i-x)}.$$

Nadaraya (1964) and Watson (1964) independently introduced this estimator which exhibits the following two highlighted features. First, it is simple to implement. Second, it is a proper estimator of the conditional CDF. That is, for $\phi(X_{i+1})=\mathbb{1}_{(-\infty,y]}(X_{i+1})$ with $y\in\mathbb{R}$, the estimator is monotone in y because the kernel function is usually assumed to be nonnegative.

The local linear estimator assumes a linear approximation in (1.2), i.e., $q=1$, and uniform probabilities. The solution of (1.3) then reads

$$\widehat{m}_{ll}(x)=\frac{\sum_{i=1}^{n}w_i(x)\phi(X_{i+1})}{\sum_{i=1}^{n}w_i(x)},$$

with weights

$$w_i(x) = K_{h_n}(X_i - x)\Bigg(\sum_{k=1}^{n}(X_k - x)^2 K_{h_n}(X_k - x) \\ -(X_i - x)\sum_{k=1}^{n}(X_k - x)K_{h_n}(X_k - x)\Bigg). \tag{1.4}$$

Stone (1977) and Cleveland (1979) introduced this estimator with Fan and Gijbels (1992), Fan (1993), and others subsequently investigating it. There are numerous advantages favoring a local linear rather than a local constant fit which are not all covered in detail here. For illustrations see Hastie and Loader (1993) and for a general treatment see Fan and Gijbels (1996, pp. 60–76). One major advantage, however, is that the local linear estimator fulfills the following condition

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - x)w_i(x) = 0. \tag{1.5}$$

A direct consequence of this discrete moment condition is that the local linear estimator exhibits zero finite sample bias when estimating linear functions. For more details regarding the above condition see Fan (1993, pp. 198-199). The Nadaraya-Watson estimator on the other hand does not fulfill an equivalent condition given the weights $K_{h_n}(X_i - x)$, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - x)K_{h_n}(X_i - x) \neq 0. \tag{1.6}$$

Because of (1.6) the finite-sample bias of the Nadaraya-Watson estimator exhibits an additional term and therefore is inferior to the bias of the local linear estimator (for details see Fan and Gijbels (1996, p. 63) and the reference therein). Because the variance of both estimators are equal (see Fan and Gijbels (1996, p. 63)), the local linear estimator is considered superior to the Nadaraya-Watson estimator.

A major drawback of the local linear estimator, discussed in more detail in Yu and Jones (1998) and Hall et al. (1999), is that if $\phi(X_{i+1}) = \mathbb{1}_{(-\infty,y]}(X_{i+1})$ for some $y \in \mathbb{R}$, the estimated conditional CDF may exhibit non-distributional properties. These properties convey in a non-monotone behavior of the conditional CDF implying the estimation of a function which essentially is not a CDF.[1] This is

[1] This is a particular problem of the analysis presented in Chesney et al. (2011, pp. 257–258).

because given the above weights $w_i(x)$ in (1.4) it is not guaranteed that these are in fact all nonnegative for every location x and bandwidth h_n.

Given the above facts it is desirable to design an estimator that reproduces the superior bias properties of local linear methods while preserving the property that the Nadaraya-Watson estimator is always a distribution function. The answer to this desire is the *weighted* Nadaraya-Watson estimator. Hall and Presnell (1999) introduce the biased bootstrap procedure leading to constrained estimators such as the weighted Nadaraya-Watson estimator in a framework with independent data. In their approach the probabilities $p_i(x;\lambda_n)$ of the minimization problem given in (1.3) are used to guarantee an equivalent condition as in (1.5). This implies that at x the probability mass is not uniform but shifted around n^{-1} such that

$$\sum_{i=1}^{n} p_i(x;\lambda_n)(X_i - x)K_{h_n}(X_i - x) = 0, \tag{1.7}$$

holds. To guarantee that $p_i(x;\lambda_n)$ are indeed probabilities it is further imposed that

$$p_i(x;\lambda_n) \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} p_i(x;\lambda_n) = 1. \tag{1.8}$$

To determine the probabilities for the constrained estimator Hall and Presnell (1999) and Hall et al. (1999) propose to select these via the empirical likelihood by maximizing the empirical log-likelihood, $\sum_{i=1}^{n} \ln(p_i(x))$, subject to the above constraints (1.7) and (1.8) whereas the strict positivity constraint of the probabilities is implicitly imposed by the objective function. Let $\lambda_n(x)$ denote the Lagrange-parameter for condition (1.7) of the reduced optimization problem, then after some algebra[2] the probabilities are given by

$$p_i(x;\lambda_n) = \frac{1}{n(1 + \lambda_n(x)(X_i - x)K_{h_n}(X_i - x))}, \tag{1.9}$$

with $\lambda_n(x)$ not having a closed form solution which makes the analysis particularly difficult. Given the empirical log-likehood function and (1.9) the optimization problem, to determine the probabilities, can be simplified leading to the maximization of $L_n(x;\lambda_n) = -\sum_{i=1}^{n} \ln(1 + \lambda_n(x)(X_i - x)K_{h_n}(X_i - x))$ with respect to $\lambda_n(x)$. The

[2]For a complete derivation see, e.g., Li and Racine (2006, pp. 186–189).

first order condition reads

$$L_n'(x;\lambda_n) = \frac{1}{nh_n}\sum_{i=1}^{n}\frac{(X_i - x)K_{h_n}(X_i - x)}{1 + \lambda_n(x)(X_i - x)K_{h_n}(X_i - x)} = 0, \tag{1.10}$$

where the multiplication by $(nh_n)^{-1}$ facilitates the proofs in the appendix. Given the optimal probabilities at x, the *weighted* Nadaraya-Watson estimator is derived by minimizing (1.3), for $q = 0$, resulting in

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} p_i(x;\lambda_n)K_{h_n}(X_i - x)\phi(X_{i+1})}{\sum_{i=1}^{n} p_i(x;\lambda_n)K_{h_n}(X_i - x)}. \tag{1.11}$$

By minimizing (1.3) with $q = 1$ and applying condition (1.7) there exists an alternative derivation of the estimator.

Applications of the estimator are mainly in the realm of estimating the conditional CDF because it avoids the deficit of the local linear method for this particular case. Kato (2012) constructs an estimator of the conditional expected shortfall using the weighted Nadaraya-Watson estimator. A similar problem, as well as the estimation of the conditional Value-at-Risk, is discussed in Cai and Wang (2008). Bao et al. (2006) evaluate the predictive performance of the estimator in Value-at-Risk models. Tay and Ting (2008) investigate the CDF of high-frequency price changes conditional on trading volume and duration between trades. Cai (2002) proposes a quantile regression estimators based on the weighted Nadaraya-Watson estimator. Steikert (2014) uses the estimator for a local bootstrap procedure to select the bandwidth for the weighted Nadaraya-Watson estimator.

1.3 Main results

The next subsection introduces and discusses the assumptions in detail. The results leading to pointwise strong consistency follow. All proofs are provided in Appendix A.

1.3.1 Assumptions

The underlying stochastic sequence $\{X_i\}_{i=1}^{n+1}$ satisfies the following sets of assumptions.

Assumptions 1.3.1.

a) $\{X_i\}_{i=1}^{n+1}$ is a sequence of strictly stationary random numbers.

b) The response function, $\phi(\cdot)$, is Borel measurable on the real line and satisfies $\mathbb{E}|\phi(X_{i+1})|^s < \infty$ for some $s > 2$.

c) The regression function $m(\cdot)$ has continuous derivatives in the neighborhood of x up to the second order.

d) At x the conditional variance $\sigma^2(x) = \text{var}(\phi(X_{i+1})|X_i = x)$ is continuous.

All of the above assumptions are commonly used in the literature. Assumption 1.3.1.a) simplifies most arguments in the proofs. Weakening stationarity is possible but comes at a notational expense without gaining further insights. See Kristensen (2009) for a uniform consistency result for kernel estimators with heterogenous data. Assuming the existence of the moments $\mathbb{E}|\phi(X_{i+1})|^s$ for some $s > 2$, where s is model dependent, is a fairly weak condition. An alternative assumption is to bound the response values which is unnecessary strong. Continuity, i.e., Assumption 1.3.1.c), is already assumed partly via the model set-up presented in Section 1.2. Weakening this assumption is possible implying a different magnitude of finite-sample bias of the estimator which in consequence would influence the convergence rate.

For the general case assume $\{(X_i, Y_i)\}_{i=1}^{n}$ to be strictly stationary. All other assumptions are similar by setting $X_{i+1} = Y_i$.

Assumptions 1.3.2.

a) The kernel function $K(\cdot)$ is a symmetric, unimodal, and bounded density, i.e, $K(u) = K(-u)$ for all $u \in \mathbb{R}$, $\exists! u \in \mathbb{R} : K(u) > K(v)$ for all $v \in \mathbb{R}$, and $\sup_{u\in\mathbb{R}} K(u) \leq C_1 < \infty$.

b) $K(\cdot)$ has compact support, i.e., $K(u) = 0$ for $|u| \geq 1$.

Assuming a symmetric and bounded kernel function with bounded support is common (see, e.g., Pagan and Ullah (1999, p. 109)). The choice of the support of the kernel function is without loss of generality, i.e., for any other strictly positive number the statements below hold. The bounded support helps in shortening the arguments of the proofs but can be removed at the expense of lengthier arguments. Assuming a single mode, however, is rather uncommon but not very restrictive because, e.g., bimodal kernels can increase the mean squared error of nonparametric

estimators. Bounded support of $K(\cdot)$ excludes kernel functions such as the Gaussian kernel. However, as presented in Fan et al. (1995), the Epanechnikov kernel function, which fulfills all the above assumptions, is optimal in a minimax sense for the local linear estimator. Further implications of Assumption 1.3.2 are the following. Lemma B.1 in Appendix B establishes $C_2 = \sup_{u\in\mathbb{R}} uK(u) < 1$. In addition, Lemmas B.2 and B.3 establish $|u|K(u) \leq C_2 < \infty$ and $\int |u|K(u)du \leq C_3 < \infty$, respectively. Given Assumption 1.3.1.c) and Assumption 1.3.2.b), all terms of the approximation in (1.2) of order greater than two are $o(h_n^2)$. This is because expanding $m(X_i)$ in the neighborhood of $X_i \in \mathcal{D}(x) = \{r \in \mathbb{R} : x - h_n \leq r \leq x + h_n\}$, i.e., around x leads to $|X_i - x| \leq h_n$. This also implies that $|m(X_i)|$ is bounded in the neighborhood of x and therefore $\sup_{X_i\in\mathcal{D}(x)} |m(X_i)| \leq C_4 < \infty$ (see Lemma B.4).

Assumptions 1.3.3.

a) For fixed x the marginal density of X_i, $f_{X_i}(x)$, is bounded away from zero and continuously differentiable up to the first order in the neighborhood of x.

b) The joint density of X_1 and X_i, for $i > 1$, denoted by $f_{X_1,X_i}(x_1, x_i)$, is bounded in the neighborhood of x, i.e., $\sup_{(x_1,x_i)\in\mathbb{R}^2} |f_{X_1,X_i}(x_1, x_i)| \leq C_5$ for $x_1, x_i \in \mathcal{D}(x)$.

c) The conditional CDF of X_{i+1} given $X_i = u$ is continuous at $u = x$.

Assumption 1.3.3.a) ensures that certain expressions are well-defined at the location x. Differentiability of $f_{X_i}(x)$ implies faster rates of convergence. Assumption 1.3.3.b) is necessary to determine asymptotic bounds for various covariance terms emerging later in the manuscript.

Because the processes under consideration are assumed to be weakly dependent let $\mathcal{B}_s^t = \sigma\{X_i : s \leq i \leq t\}$ denote the σ-algebra generated by the time series segment $\{X_s, X_{s+1}, ..., X_t\}$. Then, the α-mixing coefficient, introduced in Rosenblatt (1956), is defined as

$$\alpha(k) = \sup_{A\in\mathcal{B}_{-\infty}^0, B\in\mathcal{B}_k^\infty} \big|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)\big|, \tag{1.12}$$

assuming strict stationarity. The mixing coefficient is the total variation distance between the joint distribution and the product of the marginal distributions of the entire past and future k lags from today. It basically represents the maximum dependence between two time events at least k steps apart. The random sequence is strongly mixing if $\alpha(k) \to 0^+$ as $k \to \infty$. Strongly mixing therefore represents a

form of asymptotic independence. Regarding the α-mixing coefficient the following condition is imposed.

Assumption 1.3.4. The underlying stochastic sequence is strongly mixing with mixing coefficient $\alpha(k)$, given in (1.12), satisfying $\sum_{k=1}^{\infty} k^a \alpha^{1-2/\delta}(k) < \infty$ for some $\delta > 2$ and $a > 1 - 2/\delta$.

The strong mixing coefficient must converge to zero sufficiently fast to guarantee finiteness of the specified sum. This summability condition is needed to guarantee finite covariances for infinite sums of the underlying data $\{X_i\}_{i=1}^{n+1}$ for $n \to \infty$. To adjust Assumption 1.3.4 to the general case let $\mathcal{B}_s^t = \sigma\{(X_i, Y_i) : s \leq i \leq t\}$ and define the strong mixing coefficient similar as is (1.12).

1.3.2 Pointwise strong consistency

Because there does not exist a closed form expression for $\lambda_n(x)$, the probabilities $p_i(x; \lambda_n)$ complicate proving strong consistency for the weighted Nadaraya-Watson estimator substantially. In case of the ordinary Nadaraya-Watson estimator the proof is much simpler because as noted above there the probabilities are uniform.

In Theorem 1.3.7 below I provide the fundamental result, that asymptotically the optimal probabilities, $p_i(x; \lambda_n)$, are uniform with probability one. This implies, that with an increasing number of observations less and less probability mass is shifted around n^{-1} to guarantee conditions (1.7) and (1.8). Before constituting the theorem a series of lemmas is needed to establish the asymptotic variance and the asymptotic behavior of the partial sum $S_{n,j}(x)$ which is defined as

$$S_{n,j}(x) = \frac{1}{nh} \sum_{i=1}^{n} (X_i - x)^j K_{h_n}^j (X_i - x), \tag{1.13}$$

for integers $j > 0$. Partial sums such as $S_{n,j}(x)$, or variants thereof, commonly emerge when proving consistency for nonparametric estimators (see Fan and Gijbels (1996, p. 63) or Masry (1996b, p. 84), among others).

The following lemma establishes the asymptotic variance of (1.13). The result is related to Masry (1996b, Theorem 1) although the definitions of the partial sum $S_{n,j}(x)$ differ and Lemma 1.3.5 holds for all integers $j > 0$. It states that the asymptotic variance of (1.13) converges to zero which implies vanishing dependence for increasing sample size. While for independent data the statement is more or less obvious, here, in case of weakly dependent data, the proof is much more demanding. In particular the summation of infinite covariance terms is challenging.

Lemma 1.3.5. *Assume $h_n \to 0$, $nh_n \to \infty$, as $n \to \infty$, and that Assumptions 1.3.1–1.3.4 hold. Then at every continuity point x of $f_{X_i}(\cdot)$,*

$$nh_n var(S_{n,j}(x)) \to \nu_{2j} f_{X_i}(x), \tag{1.14}$$

for $j = 1, 2, ...,$ with $\nu_j = \int u^j K^j(u)du$.

Note that ν_{2j} is decreasing for increasing j because the kernel function $K(\cdot)$ is unimodal (see Lemma B.1). This implies, that $\nu_2 f_{X_i}(x)$ is an upper asymptotic bound for $nh_n \text{var}(S_{n,j}(x))$ for all integers $j > 0$.

Next I establish the asymptotic behavior of $S_{n,j}(x)$ by applying the Borel-Cantelli lemma. The application, however, is not straightforward because of the weakly dependent data. In the proof of Lemma 1.3.6 I make use of a coupling theorem to approximate the dependent random variables by independent ones. For the following lemma let $\lfloor x \rfloor$ denote the integer part of $x \in \mathbb{R}$.

Lemma 1.3.6. *Assume $h_n \to 0$, $nh_n/\ln(n) \to \infty$, as $n \to \infty$, and that Assumptions 1.3.1–1.3.4 hold. If*

$$\Theta_n = \left(\left(\frac{n}{h_n} \right)^3 \ln(n) \right)^{1/4} \alpha \left(\left\lfloor \sqrt{\frac{nh_n}{\ln(n)}} \right\rfloor \right),$$

is summable, i.e., $\sum_{n=1}^{\infty} \Theta_n < \infty$, then at every continuity point x of $f_{X_i}(\cdot)$,

$$S_{n,j}(x) = \begin{cases} \mu_j h_n f'_{X_i}(x) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) & \text{almost surely, for odd } j > 0; \\ \nu_j f_{X_i}(x) + \mathcal{O}(h_n^2) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) & \text{almost surely, for even } j > 0, \end{cases} \tag{1.15}$$

with $\mu_j = \int u^{j+1} K^j(u)du$ and $\nu_j = \int u^j K^j(u)du$.

The continuity assumption of the marginal density implies different results for the asymptotic behavior of the partial sum $S_{n,j}(x)$ for odd and even $j > 0$. The difference is important for proving the next result, which establishes $\lambda_n(x) \to 0$ with probability one, as $n \to \infty$.

Theorem 1.3.7. *Assume that the assumptions of Lemma 1.3.6 hold, then* $\lambda_n(x) = \mathcal{O}(h_n) + \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ *almost surely. In particular, at every continuity point* x *of* $f_{X_i}(\cdot)$,

$$\lambda_n(x) = \frac{\mu_1 h_n f'_{X_i}(x) + \mathcal{O}\Big(\sqrt{\ln(n)/(nh_n)}\Big)}{\nu_2 f_{X_i}(x)} \qquad \textit{almost surely,}$$

with $\mu_1 = \int u^2 K(u)du$ *and* $\nu_2 = \int u^2 K^2(u)du$.

Because Lemmas 1.3.5 and 1.3.6 are needed to prove Theorem 1.3.7 the stronger assumption on the bandwidth of Lemma 1.3.6, namely $nh_n/\ln(n) \to \infty$, is imposed. The implication of Theorem 1.3.7 is straightforward, asymptotically the probabilities are almost surely uniform while honoring the conditions (1.7) and (1.8).

Before proceeding to establish pointwise strong consistency for the weighted Nadaraya-Watson estimator the consistency problem is separated as follows

$$\begin{aligned}
\widehat{m}(x) - m(x) &= \frac{\sum_{i=1}^n p_i(x;\lambda_n) K_{h_n}(X_i - x)\phi(X_{i+1})}{\sum_{i=1}^n p_i(x;\lambda_n) K_{h_n}(X_i - x)} - m(x) \\
&= \frac{\sum_{i=1}^n p_i(x;\lambda_n) K_{h_n}(X_i - x)(\phi(X_{i+1}) - m(x))}{\sum_{i=1}^n p_i(x;\lambda_n) K_{h_n}(X_i - x)} \\
&= \frac{\sum_{i=1}^n p_i(x;\lambda_n) K_{h_n}(X_i - x)(\phi(X_{i+1}) - m(X_i) + m(X_i) - m(x))}{\sum_{i=1}^n p_i(x;\lambda_n) K_{h_n}(X_i - x)} \\
&= \frac{J_1(x)}{J_3(x)} + \frac{J_2(x)}{J_3(x)}, \qquad (1.16)
\end{aligned}$$

with

$$J_3(x) = \sum_{i=1}^n p_i(x;\lambda_n) K_{h_n}(X_i - x), \qquad (1.17)$$

$$J_2(x) = \sum_{i=1}^n (m(X_i) - m(x)) p_i(x;\lambda_n) K_{h_n}(X_i - x), \qquad (1.18)$$

$$J_1(x) = \sum_{i=1}^n (\phi(X_{i+1}) - m(X_i)) p_i(x;\lambda_n) K_{h_n}(X_i - x). \qquad (1.19)$$

In what follows I derive the asymptotic behavior for each of the expressions in (1.17)–(1.19) and prove that each summand in (1.16) converges to zero with probability one as $n \to \infty$. Proposition 1.3.8 states that $J_3(x)$, the sum of probability weighted kernel weights, converges to the marginal density of X_i with probability

one. Hence, asymptotically, $J_3(x)$ behaves similar to an ordinary kernel estimator of the marginal density while honoring conditions (1.7) and (1.8). Expression (1.18) is the weighted sum of the error of approximating $m(X_i)$ by $m(x)$. It is therefore related to the bias of the estimator. Proposition 1.3.9 establishes the asymptotic behavior for $J_2(x)$. Expression (1.19) is the weighted difference between the true observed value $\phi(X_{i+1})$ and $m(X_i)$ and is therefore equivalent to the weighted sum of the true model. The asymptotic behavior of $J_1(x)$ is established in Proposition 1.3.13.

Proposition 1.3.8. *Assume that the assumptions of Lemma 1.3.6 hold, then at every continuity point x of $f_{X_i}(\cdot)$,*

$$J_3(x) = f_{X_i}(x) + \mathcal{O}(h_n^2) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \textit{almost surely.}$$

Proposition 1.3.9. *Assume that the assumptions of Lemma 1.3.6 hold, then at every continuity point x of $f_{X_i}(\cdot)$,*

$$J_2(x) = \frac{1}{2}h_n^2\mu_1 f_{X_i}(x)m''(x) + o(h_n^2) \qquad \textit{almost surely,}$$

with $\mu_1 = \int u^2 K(u)du$.

To derive the asymptotic behavior for $J_1(x)$, given in (1.19), note that $\phi(X_{i+1})$ is not necessarily bounded due to Assumption 1.3.1.b). I therefore employ the well known truncation argument introduced in Mack and Silverman (1982, p. 408) to show that $|\phi(X_{i+1})|$ is almost surely bounded by τ_n, which is defined below. Let

$$T_{n,j}(x) = \frac{1}{n}\sum_{i=1}^{n}(\phi(X_{i+1}) - m(X_i))(X_i - x)^j K_{h_n}^{j+1}(X_i - x), \tag{1.20}$$

for integers $j \geq 0$. To establish the asymptotic behavior of $J_1(x)$, (1.20) serves a similar purpose as $S_{n,j}(x)$ which was needed to establish the asymptotic behavior of $\lambda_n(x)$. It is easy to see that $\mathbb{E}\,T_{n,j}(x) = 0$ because $\mathbb{E}(\epsilon_i|X_i) = 0$, where ϵ_i is the error of the underlying model $\phi(X_{i+1}) = m(X_i) + \epsilon_i$. Define the truncation of (1.20) as

$$T_{n,j}^{(t)}(x) = \frac{1}{n}\sum_{i=1}^{n}\big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}} - m^{(t)}(X_i)\big)(X_i - x)^j K_{h_n}^{j+1}(X_i - x), \tag{1.21}$$

with

$$\tau_n = \big(n(\ln(\ln(n)))^2 \ln(n)\big)^{1/s}, \tag{1.22}$$

for some $s > 2$, where the truncation of $m(X_i)$ is defined as

$$m^{(t)}(X_i) = \mathbb{E}\big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}|X_i\big).$$

The bound τ_n depends on the order s of the moment $\mathbb{E}\,|\phi(X_{i+1})|^s < \infty$. For low order, e.g., for $s = 3$, τ_n is a steep function in n, whereas for $s \to \infty$ it is flat. An example for the case of infinite order s of the moment $\mathbb{E}\,|\phi(X_{i+1})|^s < \infty$ is the conditional CDF, i.e., $\phi(X_{i+1}) = \mathbb{1}_{(-\infty,y]}(X_{i+1})$ with $y \in \mathbb{R}$. To justify working with the truncated expression in (1.21) instead of (1.20) it is necessary to prove that the substitution leads to an error of small order. For this define

$$\begin{aligned} R_{n,j}(x) &= T_{n,j}(x) - T^{(t)}_{n,j}(x) \\ &= \frac{1}{n}\sum_{i=1}^{n}\big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|>\tau_n\}} \\ &\qquad - \mathbb{E}\big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|>\tau_n\}}|X_i\big)\big)(X_i - x)^j K^{j+1}_{h_n}(X_i - x). \end{aligned} \tag{1.23}$$

Lemma 1.3.10. *Assume that Assumption 1.3.1 holds, then, for $R_{n,j}(x)$ defined in* (1.23)*,*

$$R_{n,j}(x) = o(1) \qquad \textit{almost surely,}$$

for integers $j \geq 0$.

To establish the asymptotic behavior of $T_{n,j}(x)$ it suffices to establish the asymptotic behavior for $T^{(t)}_{n,j}(x)$ due to $T_{n,j}(x) = T^{(t)}_{n,j}(x) + R_{n,j}(x)$ and Lemma 1.3.10. In the following I establish results similar to Lemmas 1.3.5 and 1.3.6.

Lemma 1.3.11. *Assume $h_n \to 0$, and $nh_n \to \infty$, as $n \to \infty$, $s \geq \delta > 2$, and that Assumptions 1.3.1–1.3.4 hold, then at every continuity point x of $C(\cdot)$ and $f_{X_i}(\cdot)$,*

$$nh_n var\Big(T^{(t)}_{n,j}(x)\Big) \to C(x)f_{X_i}(x)\int u^{2j}K^{2(j+1)}(u)du, \tag{1.24}$$

for integers $j \geq 0$, with $C(x)$ defined in (A.32)*.*

For the above lemma to hold the order s of the moment condition $\mathbb{E}\,|\phi(X_{i+1})|^s < \infty$ needs to be larger than the order δ of $\mathbb{E}\,|X_i|^\delta < \infty$. In Lemma 1.3.5 this was not necessary because there strongly mixing processes were investigated and not functions thereof. Given (1.24) I establish the asymptotic behavior of $T_{n,j}^{(t)}(x)$ similar to Lemma 1.3.6.

Lemma 1.3.12. *Assume $h_n \to 0$, $nh_n/\ln(n) \to \infty$, as $n \to \infty$, $s \geq \delta > 2$, and that Assumptions 1.3.1–1.3.4 hold. If*

$$\Xi_n = \tau_n^{3/2}\left(\ln(n)\left(\frac{n}{h_n}\right)^3\right)^{1/4}\alpha\left(\left\lfloor\frac{1}{\tau_n}\sqrt{\frac{nh_n}{\ln(n)}}\right\rfloor\right),$$

is summable, i.e., $\sum_{n=1}^{\infty}\Xi_n < \infty$, with τ_n defined in (1.22), then

$$T_{n,j}^{(t)}(x) = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \textit{almost surely,}$$

for all integers $j \geq 0$.

For Lemma 1.3.12 the bandwidth h_n needs to be sufficiently slow such that $\tau_n^{-1}\sqrt{nh_n/\ln(n)} \to \infty$. Given Lemmas 1.3.10 and 1.3.12 it is easy to see that $T_{n,j}(x) = \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ with probability one which is essential to establish the asymptotic behavior for (1.19).

Proposition 1.3.13. *Assume that the assumptions of Lemma 1.3.12 hold, then*

$$J_1(x) = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \textit{almost surely.}$$

Combining Propositions 1.3.8, 1.3.9, and 1.3.13 lead to the main theorem which establishes pointwise strong consistency for the weighted Nadaraya-Watson estimator, $\widehat{m}(x)$, given in (1.11).

Theorem 1.3.14. *Given the Assumptions of Lemma 1.3.12, then at every continuity point x of $m(\cdot)$,*

$$\widehat{m}(x) - m(x) = \mathcal{O}(h_n^2) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \textit{almost surely.}$$

Regarding the convergence rate for optimal bandwidth h_n, i.e., the bandwidth for which the convergence is fastest, I provide the following corollary.

Corollary 1.3.15. *Assume the assumptions of Theorem 1.3.14 hold and that the bandwidth is optimal, i.e.,*

$$h_n \sim \left(\frac{\ln(n)}{n}\right)^{1/5},$$

then at every continuity point x of $m(\cdot)$,

$$\widehat{m}(x) - m(x) = \mathcal{O}\left(\frac{\ln(n)}{n}\right)^{2/5} \qquad \textit{almost surely.}$$

If the bandwidth h_n is optimal, i.e., asymptotically equivalent to $(\ln(n)/n)^{1/5}$, then the weighted Nadaraya-Watson estimator converges almost surely with rate $(\ln(n)/n)^{2/5}$ to the value being estimated. This rate is the same strong rate established in Sarda and Vieu (2000, p. 62) for the ordinary Nadaraya-Watson estimator with similar assumptions but independent data. Also Cheng (1995, p. 361) established this strong convergence rate for the ordinary Nadaraya-Watson estimator with a different set of assumptions. It is important to note that the rate in Corollary 1.3.15 is penalized by $\ln(n)$. This (slow) penalty is nonexistent for pointwise weak consistency for local polynomial estimators (see, e.g., Li and Racine (2006, p. 87)). It is conjectured that in a weak convergence setting this penalty is nonexistent also for the weighted Nadaraya-Watson estimator. In this strong setting, however, it emerges and is an artifact of the proof technique used. Due to the penalization the rate is not optimal in the sense of Stone (1982, Theorem 1). This is also true for the rates established for the Nadaraya-Watson estimator in the aforementioned references. For the convergence to be optimal the rate needs to be of order $n^{-2/5}$. The rate established in Corollary 1.3.15, however, is close.

To the best of my knowledge the optimal rate has not yet been established in similar settings for similar estimators proving pointwise strong consistency. This is because the proof technique relies on the application of Borel-Cantelli lemma together with Bernstein-type inequalities which are of exponential type.

1.4 Conclusion

In this manuscript I establish pointwise strong consistency for the weighted Nadaraya-Watson estimator in case of weakly dependent data. The estimator is designed to reproduces the superior bias properties of local linear estimator while, in case of estimating the conditional CDF, preserving the property that the Nadaraya-Watson estimator is always a distribution function. The results presented here show that for optimal bandwidth, i.e., the bandwidth asymptotically equivalent to $(\ln(n)/n)^{1/5}$, the estimator convergence almost surely with rate $(\ln(n)/n)^{2/5}$ to the value being estimated. This rate is slightly penalized but equal to existing rates for the ordinary Nadaraya-Watson estimator in case of independent data.

Acknowledgements

I thank Silvia Grätz, Michael Wolf, and Jan Wrampelmeyer for valuable comments and suggestions.

Appendices

A Proofs

To verify pointwise strong consistency I chose a direct approach by expressing the probabilities $p_i(x;\lambda_n)$, given in (1.9), by its binomial representation. In what follows $\overline{C}$ denotes a suitable generic constant taking (possibly) different values at different positions throughout the proofs. The domain of the integrals below is usually $(-\infty,\infty)$ but since the kernel function has bounded support, due to Assumption 1.3.2.b), the domain is constraint to $\mathcal{D}(x) = \{r \in \mathbb{R} : x - h_n \leq r \leq x + h_n\}$ which is mostly suppressed to keep notation parsimonious. For proving the general case one may substitute $X_{i+1} = Y_i$ in the proofs below and use the set of generalized assumptions.

A.1 Proof of Lemma 1.3.5

The variance of the partial sum $S_{n,j}(x)$ in (1.13) is separated as follows

$$\begin{aligned}
\operatorname{var}(S_{n,j}(x)) &= \frac{1}{(nh_n)^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\operatorname{cov}\Big((X_i-x)^jK_{h_n}^j(X_i-x),(X_k-x)^jK_{h_n}^j(X_k-x)\Big)\\
&= \frac{1}{nh_n^2}\operatorname{var}\Big((X_1-x)^jK_{h_n}^j(X_1-x)\Big)\\
&\quad + \frac{2}{nh_n^2}\sum_{i=2}^{n}\left(1-\frac{i-1}{n}\right)\\
&\qquad \times \operatorname{cov}\Big((X_1-x)^jK_{h_n}^j(X_1-x),(X_i-x)^jK_{h_n}^j(X_i-x)\Big)\\
&=: \Sigma_{j,1}(x)+\Sigma_{j,2}(x), \qquad \text{(A.1)}
\end{aligned}$$

where the second equations follows from stationarity. To determine the asymptotic behavior of $nh_n\Sigma_{j,1}(x)$, with $\Sigma_{j,1}(x)$ defined in (A.1), note that

$$\begin{aligned}
nh_n\Sigma_{j,1}(x) &= \frac{1}{h_n}\operatorname{var}\Big((X_1-x)^jK_{h_n}^j(X_1-x)\Big)\\
&= \frac{1}{h_n}\left(\mathbb{E}(X_1-x)^{2j}K_{h_n}^{2j}(X_1-x)-\Big(\mathbb{E}(X_1-x)^jK_{h_n}^j(X_1-x)\Big)^2\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{h_n}\int (z-x)^{2j} K_{h_n}^{2j}(z-x) f_{X_i}(z)dz \\
&\qquad - h_n\bigg(\frac{1}{h_n}\int (z-x)^{j} K_{h_n}^{j}(z-x) f_{X_i}(z)dz\bigg)^2 \qquad \text{(A.2)}\\
&= \int u^{2j} K^{2j}(u) f_{X_i}(x+uh_n)du - h_n\bigg(\int u^j K^j(u) f_{X_i}(x+uh_n)du\bigg)^2 \\
&= f_{X_i}(x)\int u^{2j}K^{2j}(u)du\big(1+\mathcal{O}(h_n^2)\big) - \mathcal{O}(h_n) \\
&= \nu_{2j} f_{X_i}(x)(1+\mathcal{O}(h_n)), \qquad \text{(A.3)}
\end{aligned}
$$

at every continuity point x of $f_{X_i}(\cdot)$. The fourth equation follows from a change of variable, i.e., $u = (z-x)/h_n$. Because the kernel function has bounded support and the marginal density is continuously differentiable up to the first order in the neighborhood of x, the second to last line uses a Taylor approximation of $f_{X_i}(x+uh_n)$ at x, i.e., $f_{X_i}(x+uh_n) = f_{X_i}(x) + uh_n f'_{X_i}(x) + \mathcal{O}(h_n^2)$. Furthermore, for odd moments the kernel function is zero due to Assumption 1.3.2. To prove that (A.3) is bounded by $\nu_2 f_{X_i}(x)(1+\mathcal{O}(h_n))$ for all integers $j > 0$ note that $\nu_j = \int u^j K^j(u)du \leq C_2^{j-1}C_3$. Since $C_2 < 1$ (see Lemma B.1) it is easy to see that $nh_n\Sigma_{j,1}(x) \leq \nu_2 f_{X_i}(x)(1+\mathcal{O}(h_n))$ for all integers $j > 0$.

To establish (A.3) one can also use Bochner's lemma given in Lemma B.5 in Appendix B. If so, define $H(u) = u^{2j}K^{2j}(u)$ for the first term of (A.2), then, it easy to see that, $H(\cdot)$ fulfills conditions i)–iii) of the lemma. Moreover, define $v = z - x$ and let $g(x+v) = f_{X_i}(x+v)$, then, at every continuity point x of $f_{X_i}(\cdot)$, $h_n^{-1}\int (v/h_n)^{2j}K^{2j}(v/h_n)f_{X_i}(x+v)dv \to f_{X_i}(x)\int v^{2j}K^j(v)dv$ as $n \to \infty$. Combing this with similar derivations for the second term of (A.2) lead to $nh_n\Sigma_{j,1}(x) \to \nu_{2j}f_{X_i}(x)$ using Bochner's lemma.

To bound $nh_n\Sigma_{j,2}(x)$ asymptotically, with $\Sigma_{j,2}(x)$ defined in (A.1), note that

$$
\begin{aligned}
nh_n\Sigma_{j,2}(x) &= \frac{2}{nh_n^2}\sum_{i=2}^{n}\bigg(1-\frac{i-1}{n}\bigg) \\
&\qquad \times \operatorname{cov}\Big((X_1-x)^j K_{h_n}^j(X_1-x), (X_i-x)^j K_{h_n}^j(X_i-x)\Big) \\
&\leq \frac{2}{h_n}\sum_{i=2}^{n}\bigg(1-\frac{i-1}{n}\bigg)\Big|\operatorname{cov}\Big((X_1-x)^j K_{h_n}^j(X_1-x), (X_i-x)^j K_{h_n}^j(X_i-x)\Big)\Big| \\
&\leq \frac{2}{h_n}\sum_{i=2}^{n}\Big|\operatorname{cov}\Big((X_1-x)^j K_{h_n}^j(X_1-x), (X_i-x)^j K_{h_n}^j(X_i-x)\Big)\Big|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{2}{h_n}\sum_{i=2}^{\lfloor d_n \rfloor}\left|\operatorname{cov}\Big((X_1-x)^j K_{h_n}^j(X_1-x),(X_i-x)^j K_{h_n}^j(X_i-x)\Big)\right| \\
&\quad + \frac{2}{h_n}\sum_{i=\lfloor d_n \rfloor+1}^{\infty}\left|\operatorname{cov}\Big((X_1-x)^j K_{h_n}^j(X_1-x),(X_i-x)^j K_{h_n}^j(X_i-x)\Big)\right| \\
&=: \Sigma_{j,21}(x)+\Sigma_{j,22}(x). \qquad \text{(A.4)}
\end{aligned}
$$

The last inequality is due to the separation of the sum of absolute covariance in short and long lag length contributions. Let $\lfloor d_n \rfloor$ be the integer part of d_n with d_n fulfilling $d_n \to \infty$ and $d_n h_n \to 0$ as $n \to \infty$. Then, for each summand of $\Sigma_{j,21}(x)$, it follows that

$$
\begin{aligned}
&\frac{1}{h_n}\left|\operatorname{cov}\Big((X_1-x)^j K_{h_n}^j(X_1-x),(X_i-x)^j K_{h_n}^j(X_i-x)\Big)\right| \\
&= \frac{1}{h_n}\left|\,\mathbb{E}(X_1-x)^j K_{h_n}^j(X_1-x)(X_i-x)^j K_{h_n}^j(X_i-x) - \Big(\mathbb{E}(X_1-x)^j K_{h_n}^j(X_1-x)\Big)^2\right| \\
&\leq \frac{1}{h_n}\left|\,\mathbb{E}(X_1-x)^j K_{h_n}^j(X_1-x)(X_i-x)^j K_{h_n}^j(X_i-x)\right| + \mathcal{O}(h_n) \\
&= h_n\left|\int u^j K^j(u) v^j K^j(v) f_{X_1,X_i}(x+uh_n, x+vh_n)\,du\,dv\right| + \mathcal{O}(h_n) \\
&\leq C_5 h_n \int |u|^j K^j(u)du \int |v|^j K^j(v)dv + \mathcal{O}(h_n) \\
&= C_5 h_n \left(\int |u|^j K^j(u)du\right)^2 + \mathcal{O}(h_n) \\
&= \mathcal{O}(h_n).
\end{aligned}
$$

The second term in the second line is equivalent to the second term of (A.2). The fourth and fifth line follows from changes in variables and Assumption 1.3.3.b), respectively. The expression in the last line is bounded because $\int |u|^j K^j(u)du$ is bounded (a consequence from Lemma B.3). Thus,

$$
\begin{aligned}
\Sigma_{j,21}(x) &= \frac{2}{h_n}\sum_{i=2}^{\lfloor d_n \rfloor}\left|\operatorname{cov}\Big((X_1-x)^j K_{h_n}^j(X_1-x),(X_i-x)^j K_{h_n}^j(X_i-x)\Big)\right| \\
&\leq 2C_5 \lfloor d_n \rfloor h_n \left(\int |u|^j K^j(u)du\right)^2 \\
&\leq \overline{C} d_n h_n \\
&= \mathcal{O}(d_n h_n), \qquad \text{(A.5)}
\end{aligned}
$$

with $\Sigma_{j,21}(x) \to 0$ because $d_n h_n \to 0$ as postulated above. The second line follows from the fact that $\lfloor d_n \rfloor \leq d_n$ for all n.

To bound $\Sigma_{j,22}(x)$ of (A.4) each summand of $\Sigma_{j,22}(x)$ is bounded separately using Davydov's lemma, given in Lemma B.6 in Appendix B. To apply the lemma it suffices to prove $(\mathbb{E}\,|(X_1-x)^j K_{h_n}^j(X_1-x)|^\delta)^{1/\delta} < \infty$, for some $\delta > 2$, due to stationarity. Thus,

$$\begin{aligned}
&\Big(\mathbb{E}\,\big|(X_1-x)^j K_{h_n}^j(X_1-x)\big|^\delta\Big)^{1/\delta} \\
&\qquad \le \left(\int \left|\frac{z-x}{h_n}\right|^{\delta j} K^{\delta j}\left(\frac{z-x}{h_n}\right) f_{X_i}(z)dz\right)^{1/\delta} \\
&\qquad = h_n^{1/\delta}\left(\int |u|^{\delta j} K^{\delta j}(u) f_{X_i}(x+uh_n)du\right)^{1/\delta} \\
&\qquad = h_n^{1/\delta}\left(\int |u|^{\delta j} K^{\delta j}(u)\big(f_{X_i}(x) + uh_n f'_{X_i}(x) + \mathcal{O}(h_n^2)\big)du\right)^{1/\delta} \\
&\qquad = h_n^{1/\delta}\overline{C}(1+\mathcal{O}(h_n)) \\
&\qquad < \infty,
\end{aligned}$$

because of the bounded support of the kernel function. Hence, by virtue of Davydov's lemma each summand of $\Sigma_{j,22}(x)$ is bounded, i.e.,

$$\begin{aligned}
&\frac{2}{h_n}\Big|\mathrm{cov}\Big((X_1-x)^j K_{h_n}^j(X_1-x),(X_i-x)^j K_{h_n}^j(X_i-x)\Big)\Big| \\
&\qquad\qquad \le \frac{\overline{C}(1+\mathcal{O}(h_n))}{h_n^{1-2/\delta}}(\alpha(i-1))^{1-2/\delta}.
\end{aligned}$$

Considering the sum $\Sigma_{j,22}(x)$ note that summability is of only interest, thus it suffices to show that, for $k = i-1$ and $a > 1-2/\delta$,

$$\begin{aligned}
\Sigma_{j,22}(x) &\le \frac{\overline{C}}{h_n^{1-2/\delta}} \sum_{k=\lfloor d_n\rfloor}^{\infty} (\alpha(k))^{1-2/\delta} \\
&\le \frac{\overline{C}}{h_n^{1-2/\delta}} \sum_{k=\lfloor d_n\rfloor}^{\infty} \left(\frac{k}{\lfloor d_n\rfloor}\right)^a (\alpha(k))^{1-2/\delta} \\
&= \frac{\overline{C}}{h_n^{1-2/\delta}\lfloor d_n\rfloor^a} \sum_{k=\lfloor d_n\rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta} \\
&= \overline{C} \sum_{k=\lfloor d_n\rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta} \\
&= \mathrm{o}(1). \qquad (A.6)
\end{aligned}$$

The second line follows from the fact that $(k/\lfloor d_n\rfloor)^a \ge 1$ for integers $k \ge \lfloor d_n\rfloor$ and any $a > 0$. For the second to last line define $d_n = h_n^{-(1-2/\delta)/a}$, then, as $n \to \infty$, $d_n \to \infty$

and $d_n h_n \to 0$ because $(1 - 2/\delta)/a \in (0, 1)$. Moreover, note that $\lfloor d_n \rfloor^a = (d_n - z)^a$ with $z \in [0, 1)$ and therefore $\lfloor d_n \rfloor^a$ and d_n^a are asymptotically equivalent. Finally, the contributions of the sum $\sum_{k=\lfloor d_n \rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta}$ must fade to zero because $d_n \to \infty$ as $n \to \infty$ and the summability condition of Assumption 1.3.4. Combining (A.4), (A.5), and (A.6) proves

$$n h_n \Sigma_{j,2}(x) \to 0, \tag{A.7}$$

as $n \to \infty$. Thus, (A.1), (A.3), and (A.7) proves the statement of the lemma.

□

A.2 Proof of Lemma 1.3.6

To prove the statement note that,

$$\big|S_{n,j}(x) - \mu_j h_n f'_{X_i}(x)\big| \leq \big|\,\mathbb{E}\, S_{n,j}(x) - \mu_j h_n f'_{X_i}(x)\big| + \big|S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x)\big|, \tag{A.8}$$

for odd $j > 0$ and

$$\big|S_{n,j}(x) - \nu_j f_{X_i}(x)\big| \leq \big|\,\mathbb{E}\, S_{n,j}(x) - \nu_j f_{X_i}(x)\big| + \big|S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x)\big|, \tag{A.9}$$

for even $j > 0$. Establishing the first terms of the right side of (1.15) for both inequalities is simple and follows from

$$\begin{aligned}
\mathbb{E}\, S_{n,j}(x) &= \frac{1}{h_n}\,\mathbb{E}(X_1 - x)^j K_{h_n}^j(X_1 - x) \\
&= \frac{1}{h_n}\int (z - x)^j K_{h_n}^j(z - x) f_{X_i}(z) dz \\
&= \int u^j K^j(u)\big(f_{X_i}(x) + u h_n f'_{X_i}(x) + \mathcal{O}(h_n^2)\big) du \\
&= \begin{cases} \mu_j h_n f'_{X_i}(x) + \mathrm{o}(h_n^2), & \text{for odd } j > 0; \\ \nu_j f_{X_i}(x) + \mathcal{O}(h_n^2), & \text{for even } j > 0, \end{cases}
\end{aligned}$$

at every continuity point x of $f_{X_i}(\cdot)$. The first line follows from stationarity, the third from a change of variable ($u = (z - x)/h_n$) and a Taylor expansion of the marginal density $f_{X_i}(x + u h_n)$ at x, and the last from $K(\cdot)$ being a symmetric kernel function implying zero odd moments. Given the above derivations it is easy to see that

$$\big|\,\mathbb{E}\, S_{n,j}(x) - \mu_j h_n f'_{X_i}(x)\big| = \mathrm{o}(h_n^2), \quad \text{for odd } j > 0, \tag{A.10}$$

and

$$\big|\,\mathbb{E}\,S_{n,j}(x) - \nu_j f_{X_i}(x)\big| = \mathcal{O}(h_n^2), \quad \text{for even } j > 0. \tag{A.11}$$

To prove $\mathbb{E}\,S_{n,j}(x) \to \nu_j f_{X_i}(x)$ as $n \to \infty$ it is sometimes reasonable to use Bochner's lemma, given in Lemma B.5. If so, define $H(v/h_n) = (v/h_n)^j K^j(v/h_n)$, with $v = z - x$, and

$$g_n(x) = \frac{1}{h_n}\int H\left(\frac{v}{h_n}\right) f(x+v)dv.$$

The approach, however, fails to provide rates of convergence which are necessary for the future analysis.

To prove $|S_{n,j}(x) - \mathbb{E}\,S_{n,j}(x)| = \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ with probability one I employ the Borel-Cantelli lemma. Define the random variable $Z_{n,j,i}(x)$ such that

$$Z_{n,j,i}(x) = \frac{1}{h_n}\Big((X_i - x)^j K_{h_n}^j(X_i - x) - \mathbb{E}(X_1 - x)^j K_{h_n}^j(X_1 - x)\Big), \tag{A.12}$$

and note that $\mathbb{E}\,Z_{n,j,i}(x) = 0$. If $\sum_{n=1}^{\infty}\mathbb{P}\big(n^{-1}\big|\sum_{i=1}^{n} Z_{n,j,i}(x)\big| > \varepsilon\big) < \infty$, then, by virtue of the Borel-Cantelli lemma, the statement follows with the convergence rate depending on ε. Thus, the task below is to prove summability of $\mathbb{P}\big(n^{-1}\big|\sum_{i=1}^{n} Z_{n,j,i}(x)\big| > \varepsilon\big)$. Note that $Z_{n,j,i}(x)$ are not independent, however, they can be approximated by independent random variables using the coupling theorem by Bradley (1983, Theorem 3) (see Lemma B.9 in Appendix B). Using the theorem to prove consistency is commonly used in the literature (see for example Tran (1990, Theorem 2.1), Masry (1996a, Theorem 1), or Lu and Cheng (1997, Theorems 2.1–2.4), among others).

To apply Lemma B.9 consider a partition of the set $\{1, 2, ..., n\}$ into $2q_n$ consecutive blocks with each block containing s_n elements. If n is odd, then there exist $2q_n + 1$ blocks with the last block containing strictly less than s_n elements. Define

$$V_{n,j,k}(x) = \frac{1}{n}\sum_{i=(k-1)s_n+1}^{ks_n} Z_{n,j,i}(x), \tag{A.13}$$

for $k = 1, ..., 2q_n$, then

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} Z_{n,j,i}(x) &= \sum_{k=1}^{2q_n} V_{n,j,k}(x) + \frac{1}{n}\sum_{i=2q_n s_n+1}^{n} Z_{n,j,i}(x) \\
&= \sum_{k=1}^{q_n} V_{n,j,2k-1}(x) + \sum_{k=1}^{q_n} V_{n,j,2k}(x) + \frac{1}{n}\sum_{i=2q_n s_n+1}^{n} Z_{n,j,i}(x) \\
&=: W'_{n,j}(x) + W''_{n,j}(x) + W'''_{n,j}(x),
\end{aligned}$$

where $W'_{n,j}(x)$ and $W''_{n,j}(x)$ contain odd and even numbered blocks, respectively. Thus, the statement follows if for every $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_{n,j,i}(x)\right| > \varepsilon\right) &= \mathbb{P}\big(\big|W'_{n,j}(x) + W''_{n,j}(x) + W'''_{n,j}(x)\big| > \varepsilon\big) \\ &\leq \mathbb{P}\Big(\big|W'_{n,j}(x)\big| > \frac{\varepsilon}{3}\Big) + \mathbb{P}\Big(\big|W''_{n,j}(x)\big| > \frac{\varepsilon}{3}\Big) \\ &\quad + \mathbb{P}\Big(\big|W'''_{n,j}(x)\big| > \frac{\varepsilon}{3}\Big),\end{aligned} \tag{A.14}$$

is summable. To bound $\mathbb{P}(|W'_{n,j}(x)| > \eta)$, with $\eta > 0$, approximate $V_{n,j,1}(x), V_{n,j,3}(x), ..., V_{n,j,2q_n-1}(x)$ by independent random variables in the following way. Introduce a sequence of independent standard uniform random variables $U_1, U_2, ...,$ independent of $V_{n,j,1}(x), V_{n,j,3}(x), ..., V_{n,j,2q_n-1}(x)$, by enlarging the probability space if necessary. Define $V^*_{n,j,0}(x) = 0, V^*_{n,j,1}(x) = V_{n,j,1}(x)$, then, due to Bradley (1983), for $k \geq 2$ there exists a real valued random variable $V^*_{n,j,2k-1}(x)$, which is a measurable function of $V^*_{n,j,1}(x), V^*_{n,j,3}(x)$, $..., V^*_{n,j,2k-3}(x)$ as well as of $V_{n,j,2k-1}(x)$ and U_k such that

i) $V^*_{n,j,2k-1}(x)$ is independent of $V^*_{n,j,1}(x), V^*_{n,j,3}(x), ..., V^*_{n,j,2k-3}(x)$,

ii) $V^*_{n,j,2k-1}(x)$ and $V_{n,j,2k-1}(x)$ have the same distribution,

iii) for $\eta \in (0, \|V_{n,j,2k-1}(x)\|_\infty]$,[3]

$$\begin{aligned}&\mathbb{P}\big(\big|V^*_{n,j,2k-1}(x) - V_{n,j,2k-1}(x)\big| > \eta\big) \\ &\quad \leq 18\sqrt{\frac{\|V_{n,j,2k-1}(x)\|_\infty}{\eta}}\left(\sup_{\substack{A\in\sigma\{V_{n,j,2t-1}(x):1\leq t\leq k-1\}\\ B\in\sigma\{V_{n,j,2k-1}(x)\}}} \big|\,\mathbb{P}(A\cap B) - \mathbb{P}(A)\,\mathbb{P}(B)\big|\right).\end{aligned}$$

In the following I refer to the above statements by simply stating its numeration i), ii), or iii). Note that the above approximation is different from the one usually used due to

[3]With $\|V(x)\|_\infty = \inf\{c \in [0,\infty] : |V(x)| \leq c \text{ almost surely}\}$.

Lu and Cheng (1997, Remark 3.1). It follows that,

$$\begin{aligned}
\mathbb{P}\Big(\big|W'_{n,j}(x)\big| > \frac{\varepsilon}{3}\Big) &= \mathbb{P}\left(\left|\sum_{k=1}^{q_n} V_{n,j,2k-1}(x)\right| > \frac{\varepsilon}{3}\right) \\
&= \mathbb{P}\left(\left|\sum_{k=1}^{q_n} \big(V_{n,j,2k-1}(x) - V^*_{n,j,2k-1}(x) + V^*_{n,j,2k-1}(x)\big)\right| > \frac{\varepsilon}{3}\right) \\
&\le \mathbb{P}\left(\left|\sum_{k=1}^{q_n} V^*_{n,j,2k-1}(x)\right| > \frac{\varepsilon}{6}\right) \\
&\quad + \mathbb{P}\left(\left|\sum_{k=1}^{q_n} V^*_{n,j,2k-1}(x) - V_{n,j,2k-1}(x)\right| > \frac{\varepsilon}{6}\right).
\end{aligned} \tag{A.15}$$

The first term of (A.15) is bounded below by Bernstein's inequality given in Lemma B.8 of Appendix B because of i). To apply the lemma the almost sure bound b of $|V^*_{n,j,2k-1}(x)|$ and $\sum_{k=1}^{q_n} \mathbb{E}(V^*_{n,j,2k-1}(x))^2$ need to be determined. To determine the bound b note that

$$\begin{aligned}
\big|Z_{n,j,i}(x)\big| &= \frac{1}{h_n}\big|(X_i - x)^j K^j_{h_n}(X_i - x) - \mathbb{E}(X_1 - x)^j K^j_{h_n}(X_1 - x)\big| \\
&\le \frac{1}{h_n}\Big(\big|(X_i - x)^j K^j_{h_n}(X_i - x)\big| + \mathbb{E}\,\big|(X_1 - x)^j K^j_{h_n}(X_1 - x)\big|\Big) \\
&\le \frac{2C_2^j}{h_n} \\
&\le \frac{2C_2}{h_n},
\end{aligned}$$

for all integers $j > 0$ since $|u|K(u) \le C_2$. The last line follows from $C_2 < 1$ due to Lemma B.1. Thus,

$$\big|V_{n,j,2k-1}(x)\big| = \frac{1}{n}\left|\sum_{i=2(k-1)s_n+1}^{(2k-1)s_n} Z_{n,j,i}(x)\right| \le \frac{2C_2 s_n}{nh_n}.$$

Because of ii), $V^*_{n,j,2k-1}(x)$ and $V_{n,j,2k-1}(x)$ have the same bound, i.e.,

$$\big|V^*_{n,j,2k-1}(x)\big| \le \frac{2C_2 s_n}{nh_n} =: b. \tag{A.16}$$

To determine $\sum_{k=1}^{q_n} \mathbb{E}(V^*_{n,j,2k-1}(x))^2$ note that, due to ii) and stationarity,

$$\begin{aligned}
\sum_{k=1}^{q_n} \mathbb{E}\big(V^*_{n,j,2k-1}(x)\big)^2 &= \sum_{k=1}^{q_n} \mathbb{E}(V_{n,j,2k-1}(x))^2 \\
&= \frac{1}{n^2} \sum_{k=1}^{q_n} \mathbb{E}\left(\sum_{i=2(k-1)s_n+1}^{(2k-1)s_n} Z_{n,j,i}(x) \right)^2 \\
&= \frac{1}{n^2} \sum_{k=1}^{q_n} \left(\sum_{i=2(k-1)s_n+1}^{(2k-1)s_n} \sum_{l=2(k-1)s_n+1}^{(2k-1)s_n} \operatorname{cov}(Z_{n,j,i}(x), Z_{n,j,l}(x)) \right) \\
&\leq \frac{1}{n^2} \sum_{i=1}^{n} \sum_{l=1}^{n} \big|\operatorname{cov}(Z_{n,j,i}(x), Z_{n,j,l}(x))\big| \\
&= \frac{1}{n} \left(\operatorname{var}(Z_{n,j,1}(x)) + 2 \sum_{i=2}^{n} \left(1 - \frac{i-1}{n}\right) \big|\operatorname{cov}(Z_{n,j,1}(x), Z_{n,j,i}(x))\big| \right) \\
&\leq \frac{1}{nh_n} \nu_{2j} f_{X_i}(x)(1 + \mathrm{o}(1)). \qquad \text{(A.17)}
\end{aligned}$$

The last line follows from the proof of Lemma 1.3.5 because for the variance term of the second to last line it follows that

$$\begin{aligned}
\frac{1}{n}\operatorname{var}(Z_{n,j,1}(x)) &= \frac{1}{nh_n^2} \operatorname{var}\big((X_i - x)^j K_{h_n}(X_i - x) - \mathbb{E}(X_1 - x)^j K_{h_n}(X_1 - x)\big) \\
&= \frac{1}{nh_n^2} \mathbb{E}\big((X_i - x)^j K_{h_n}(X_i - x) - \mathbb{E}(X_1 - x)^j K_{h_n}(X_1 - x)\big)^2 \\
&= \frac{1}{nh_n^2} \operatorname{var}\Big((X_1 - x)^j K^j_{h_n}(X_1 - x)\Big) \\
&= \Sigma_{j,1}(x) \\
&= \frac{1}{nh_n} \nu_{2j} f_{X_i}(x)(1 + \mathcal{O}(h_n)),
\end{aligned}$$

due to (A.3) leading to the first term of (A.17). A similar derivation holds for the vanishing covariance term resulting in (A.17). Using Bernstein's inequality together with (A.16), (A.17), and choosing $\varepsilon = \varepsilon_n = C_\varepsilon \sqrt{\ln(n)/(nh_n)}$, with $C_\varepsilon > 0$, and $s_n = \lfloor \sqrt{nh_n / \ln(n)} \rfloor$, the following bound for the first term of (A.15) emerges,

$$\begin{aligned}
\mathbb{P}\left(\left| \sum_{k=1}^{q_n} V^*_{n,j,2k-1}(x) \right| > \frac{\varepsilon_n}{6} \right) &\leq 2 \exp\left\{ - \frac{\frac{1}{2}\frac{1}{36} C_\varepsilon^2 \frac{\ln(n)}{nh_n}}{\frac{\nu_{2j} f_{X_i}(x)}{nh_n} + \frac{1}{3}\frac{1}{6} C_2 \frac{2s_n}{nh_n} C_\varepsilon \sqrt{\frac{\ln(n)}{nh_n}}} \right\} \\
&\leq 2 \exp\left\{ - \frac{C_\varepsilon^2 \ln(n)}{72\big(\nu_2 f_{X_i}(x) + \frac{1}{9} C_2 C_\varepsilon\big)} \right\} \\
&= 2n^{-\beta}, \qquad \text{(A.18)}
\end{aligned}$$

with $\beta = C_\varepsilon^2/(72(\nu_2 f_{X_i}(x) + 4C_2C_\varepsilon/9))$. For the second inequality use the fact that ν_{2j} is decreasing in integers $j > 0$ and that $s_n \leq \sqrt{nh_n/\ln(n)}$.

To bound the second term of (A.15) note that

$$\mathbb{P}\left(\left|\sum_{k=1}^{q_n} V^*_{n,j,2k-1}(x) - V_{n,j,2k-1}(x)\right| > \frac{\varepsilon}{6}\right) \leq \sum_{k=1}^{q_n} \mathbb{P}\left(\left|V^*_{n,j,2k-1}(x) - V_{n,j,2k-1}(x)\right| > \frac{\varepsilon}{6q_n}\right), \quad \text{(A.19)}$$

and use iii) to bound each summand of (A.19). However, since $\eta \in (0, \|V_{n,j,2k-1}(x)\|_\infty]$ it is not clear if $\varepsilon/(6q_n) < \|V_{n,j,2k-1}(x)\|_\infty$. Thus,

$$\begin{aligned}
&\mathbb{P}\left(\left|V^*_{n,j,2k-1}(x) - V_{n,j,2k-1}(x)\right| > \min\left\{\frac{\varepsilon}{6q_n}, \|V_{n,j,2k-1}(x)\|_\infty\right\}\right) \\
&\quad \leq 18\sqrt{\frac{\|V_{n,j,2k-1}(x)\|_\infty}{\min\left\{\frac{\varepsilon}{6q_n}, \|V_{n,j,2k-1}(x)\|_\infty\right\}}}\left(\sup_{\substack{A\in\sigma\{V_{n,j,2t-1}(x):1\leq t\leq k-1\}\\ B\in\sigma\{V_{n,j,2k-1}(x)\}}} \left|\mathbb{P}(A\cap B) - \mathbb{P}(A)\,\mathbb{P}(B)\right|\right) \\
&\quad \leq 18\sqrt{\frac{\|V_{n,j,2k-1}(x)\|_\infty}{\min\left\{\frac{\varepsilon}{6q_n}, \|V_{n,j,2k-1}(x)\|_\infty\right\}}}\alpha(s_n+1) \\
&\quad = 18\sqrt{\max\left\{\frac{6q_n\|V_{n,j,2k-1}(x)\|_\infty}{\varepsilon}, 1\right\}}\alpha(s_n+1) \\
&\quad \leq 18\sqrt{\max\left\{\frac{12C_2q_ns_n}{nh_n\varepsilon}, 1\right\}}\alpha(s_n+1) \\
&\quad \leq \overline{C}\sqrt{\max\left\{\frac{1}{h_n\varepsilon}, 1\right\}}\alpha(s_n+1) \\
&\quad = \overline{C}\sqrt{\frac{1}{h_n\varepsilon}}\alpha(s_n+1), \qquad \text{(A.20)}
\end{aligned}$$

for sufficiently large n and therefore

$$\mathbb{P}\left(\left|V^*_{n,j,2k-1}(x) - V_{n,j,2k-1}(x)\right| > \frac{\varepsilon}{6q_n}\right) \leq \overline{C}\sqrt{\frac{1}{h_n\varepsilon}}\alpha(s_n+1). \qquad \text{(A.21)}$$

To derive (A.20) remember that $\|V_{n,j,2k-1}(x)\|_\infty \leq 2C_2s_n/(nh_n)$ because $|V_{n,j,2k-1}(x)| \leq 2C_2s_n/(nh_n)$. The second to last line follows from $q_ns_n \leq n$ and the last line is due to the fact that for sufficiently large n, $(h_n\varepsilon)^{-1} > 1$ since $h_n \to 0$ and the fact that ε is small and positive. For the probabilities of the second line note that $\sigma\{V_{n,j,2t-1}(x) : 1 \leq t \leq k-1\} = \sigma\{X_1, X_2, ..., X_{s_n}, X_{2s_n+1}, ..., X_{3s_n}, ..., X_{(2k-3)s_n}\}$ which is a subset of $\sigma\{X_1, X_2, ..., X_{(2k-3)s_n}\}$ and $\sigma\{V_{n,j,2k-1}(x)\} = \sigma\{X_{2(k-1)s_n+1}, ..., X_{(2k-1)s_n}\}$.

Thus,

$$\sup_{\substack{A\in\sigma\{V_{n,j,2t-1}(x):1\le t\le k-1\}\\ B\in\sigma\{V_{n,j,2k-1}(x)\}}} \left|\,\mathbb{P}(A\cap B)-\mathbb{P}(A)\,\mathbb{P}(B)\right| \le \alpha(s_n+1),$$

since with $2(k-1)s_n+1-(2k-3)s_n = s_n+1$. Given (A.19), (A.21), and choosing $\varepsilon=\varepsilon_n = C_\varepsilon\sqrt{\ln(n)/(nh_n)}$ it follows that

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{k=1}^{q_n} V_{n,j,2k-1}(x) - V^*_{n,j,2k-1}(x)\right| > \frac{\varepsilon}{6}\right) &\le q_n\overline{C}\sqrt{\frac{1}{\varepsilon_n h_n}}\alpha(s_n+1)\\
&\le \frac{n\overline{C}}{s_n}\left(\frac{n}{h_n\ln(n)}\right)^{1/4}\alpha(s_n)\\
&= \overline{C}\left(\left(\frac{n}{h_n}\right)^3\ln(n)\right)^{1/4}\alpha(s_n)\\
&= \overline{C}\Theta_n, \qquad\qquad (A.22)
\end{aligned}$$

with Θ_n defined as in Lemma 1.3.6. The second inequality is due to the fact that $q_n \le n/s_n$ and $\alpha(s_n+1)\le\alpha(s_n)$ since the strong mixing coefficient is monotone.

To conclude, combine (A.15), (A.18), and (A.22) to find

$$\mathbb{P}\Big(\big|W'_{n,j}(x)\big| > \frac{\varepsilon}{3}\Big) \le 2n^{-\beta} + \overline{C}\Theta_n, \qquad (A.23)$$

for any integer $j>0$ and with $\beta = C_\varepsilon^2/(72(\nu_2 f_{X_i}(x)+4C_2C_\varepsilon/9))$. Note that β is increasing in C_ε with $\varepsilon_n\to 0$, as $n\to\infty$, for an arbitrary choice of $C_\varepsilon>0$. This guarantees the summability of the first term of (A.23). The second term is summable by assumption implying the summability of (A.23).

To prove summability of the remaining terms of (A.14) similar steps show that

$$\mathbb{P}\Big(\big|W''_{n,j}(x)\big| > \frac{\varepsilon}{3}\Big) \le 2n^{-\beta} + \overline{C}\Theta_n. \qquad (A.24)$$

The third term of (A.14) is summable because if $\mathbb{P}\Big(|W'_{n,j}(x)|>\varepsilon/3\Big)$ is summable, then $\mathbb{P}\Big(|W'''_{n,j}(x)|>\varepsilon/3\Big)$ also is because it consists of at most s_n-1 terms and therefore

$$\mathbb{P}\Big(\big|W'''_{n,j}(x)\big| > \frac{\varepsilon}{3}\Big) \le 2n^{-\beta} + \overline{C}\Theta_n. \qquad (A.25)$$

Finally, substituting the summable expressions (A.23), (A.24), and (A.25) into (A.14)

results in

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} Z_{n,j,i}(x)\right| > \varepsilon_n\right) \leq 6n^{-\beta} + \overline{C}\Theta_n, \tag{A.26}$$

which is then summable itself and therefore

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} Z_{n,j,i}(x)\right| > \varepsilon_n\right) < \infty.$$

By virtue of the Borel-Cantelli lemma, $n^{-1}|\sum_{i=1}^{n} Z_{n,j,i}(x)| = \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ almost surely and, by definition of $Z_{n,j,i}(x)$ (see (A.12)), it follows that

$$\left|S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x)\right| = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \text{almost surely,} \tag{A.27}$$

for all integers $j > 0$.

To complete the proof combine (A.8), (A.10), and (A.27) to find

$$\left|S_{n,j}(x) - \mu_j h_n f'_{X_i}(x)\right| = \mathrm{o}(h_n^2) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \text{almost surely,}$$

for odd $j > 0$, and combine (A.9), (A.11), and (A.27) to find

$$\left|S_{n,j}(x) - \nu_j f_{X_i}(x)\right| = \mathcal{O}(h_n^2) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \text{almost surely,}$$

for even $j > 0$. This completes the proof.

□

A.3 Proof of Theorem 1.3.7

The first steps of the proof are similar to Chen and Hall (1993, pp. 1174–1175). To prove the first part of the statement use the first-order condition (1.10) and note that

$$\begin{aligned}
L_n'(x;\lambda_n) &= \frac{1}{nh}\sum_{i=1}^{n}\frac{(X_i-x)K_{h_n}(X_i-x)}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)} \\
&= \frac{1}{nh}\left|\sum_{i=1}^{n}\frac{-(X_i-x)K_{h_n}(X_i-x)}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}\right| \\
&= \frac{1}{nh}\left|\sum_{i=1}^{n}\left(\frac{\lambda_n(x)((X_i-x)K_{h_n}(X_i-x))^2}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)} - (X_i-x)K_{h_n}(X_i-x)\right)\right| \\
&\geq \frac{1}{nh}\left|\sum_{i=1}^{n}\frac{\lambda_n(x)((X_i-x)K_{h_n}(X_i-x))^2}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}\right| - \left|S_{n,1}(x)\right| \\
&\geq \frac{|\lambda_n(x)|S_{n,2}(x)}{1+C_2|\lambda_n(x)|} - \left|S_{n,1}(x)\right|,
\end{aligned}$$

with $C_2 = \sup_{u\in\mathbb{R}} uK(u)$ and $S_{n,j}(x)$ defined in (1.13). The last line follows from the fact that $1+\lambda_n(x)(X_i-x)K_h(X_i-x) = |1+\lambda_n(x)(X_i-x)K_h(X_i-x)|$ because $p_i(x;\lambda_n) > 0$, thus $|1+\lambda_n(x)(X_i-x)K_h(X_i-x)| \leq 1+C_2|\lambda_n(x)|$. Because $L_n'(x;\lambda_n) = 0$, the bound for $\lambda_n(x)$ is therefore given by

$$|\lambda_n(x)| \leq \frac{|S_{n,1}(x)|}{S_{n,2}(x) - C_2|S_{n,1}(x)|}.$$

Using (1.15) of Lemma 1.3.6 it follows that

$$\begin{aligned}
\frac{1}{S_{n,2}(x) - C_2|S_{n,1}(x)|} &= \frac{1}{\nu_2 f_{X_i}(x)\Big(1+\mathcal{O}(h_n)+\mathcal{O}\Big(\sqrt{\ln(n)/(nh_n)}\Big)\Big)} \quad \text{almost surely} \\
&= \frac{1}{\nu_2 f_{X_i}(x)}\Big(1+\mathcal{O}(h_n)+\mathcal{O}\Big(\sqrt{\ln(n)/(nh_n)}\Big)\Big) \quad \text{almost surely.}
\end{aligned}$$

Thus, given $S_{n,1}(x) = \mathcal{O}(h_n) + \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ with probability one, it is easy to see that

$$\frac{|S_{n,1}(x)|}{S_{n,2}(x) - C_2|S_{n,1}(x)|} = \mathcal{O}(h_n) + \mathcal{O}\Big(\sqrt{\ln(n)/(nh_n)}\Big) \quad \text{almost surely,}$$

and therefore

$$\lambda_n(x) = \mathcal{O}(h_n) + \mathcal{O}\Big(\sqrt{\ln(n)/(nh_n)}\Big) \quad \text{almost surely,} \tag{A.28}$$

which proves the first statement of the theorem.

For the second statement note that since $\sum_{i=1}^{n} p_i(x;\lambda_n) = 1$, due to condition (1.8), it follows that

$$\begin{aligned}
1 &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=0}^{\infty}(-\lambda_n(x)(X_i-x)K_{h_n}(X_i-x))^j\right) \\
&= 1-\lambda_n(x)\frac{1}{n}\sum_{i=1}^{n}(X_i-x)K_{h_n}(X_i-x)+\lambda_n^2(x)\frac{1}{n}\sum_{i=1}^{n}(X_i-x)^2K_{h_n}^2(X_i-x) \\
&\quad -\lambda_n^3(x)\frac{1}{n}\sum_{i=1}^{n}(X_i-x)^3K_{h_n}^3(X_i-x)+\lambda_n^4(x)\frac{1}{n}\sum_{i=1}^{n}(X_i-x)^4K_{h_n}^4(X_i-x)-... \\
&= 1-h_n\big(\lambda_n(x)S_{n,1}(x)-\lambda_n^2(x)S_{n,2}(x)+\lambda_n^3(x)S_{n,3}(x)-\lambda_n^4(x)S_{n,4}(x)+...\big),
\end{aligned}$$

with $S_{n,j}(x)$ defined in (1.13). The second line follows from the binomial series representation of the probabilities $p_i(x;\lambda_n)$. Solving for $\lambda_n(x)$ leads to

$$\lambda_n(x) = \frac{S_{n,1}(x)}{S_{n,2}(x)}+\lambda_n^2(x)\frac{S_{n,3}(x)}{S_{n,2}(x)}-\lambda_n^3(x)\frac{S_{n,4}(x)}{S_{n,2}(x)}+\frac{1}{S_{n,2}(x)}\sum_{j=5}^{\infty}(-\lambda_n(x))^{j-1}S_{n,j}(x).$$

Because of (A.28) and the results of Lemma 1.3.6 it follows that

$$\frac{1}{S_{n,2}(x)}\sum_{j=3}^{\infty}(-\lambda_n(x))^{j-1}S_{n,j}(x) = \mathcal{O}\left(\left(h_n+\sqrt{\ln(n)/(nh_n)}\right)^3\right) \qquad \text{almost surely,}$$

and therefore

$$\lambda_n(x) = \frac{\mu_1 h_n f'_{X_i}(x)+\mathcal{O}\left(\sqrt{\ln(n)/(nh_n)}\right)}{\nu_2 f_{X_i}(x)} \qquad \text{almost surely,}$$

at every continuity point x of $f_{X_i}(\cdot)$. This completes the proof.

□

A.4 Proof of Proposition 1.3.8

This proof follows similar arguments as the proofs of Lemmas 1.3.5 and 1.3.6. Hence, only the necessary steps are provided. It is easy to see that

$$\begin{aligned} J_3(x) &= \sum_{i=1}^{n} p_i(x;\lambda_n) K_{h_n}(X_i - x) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=0}^{\infty}(-\lambda_n(x)(X_i-x)K_{h_n}(X_i-x))^j\right)K_{h_n}(X_i-x) \\ &= S'_{n,0}(x) - \lambda_n(x)S'_{n,1}(x) + \lambda_n^2(x)S'_{n,2}(x) - \lambda_n^3(x)S'_{n,3}(x) + \dots, \end{aligned} \tag{A.29}$$

with $S'_{n,j}(x) = n^{-1}\sum_{i=1}^{n}(X_i-x)^j K_{h_n}^{j+1}(X_i-x)$ for integers $j \geq 0$ not to be confused with, but similar to, $S_{n,j}(x)$ defined in (1.13). In what follows I establish the asymptotic behavior of $S'_{n,j}(x)$ similar to Lemma 1.3.6. To do so I first need to determine $nh_n\text{var}(S'_{n,j}(x))$ which is related to the proof of Lemma 1.3.5. For this only minor adjustments to the proof of Lemma 1.3.5 have to be made to find that, at every continuity point x of $f_{X_i}(\cdot)$,

$$nh_n\text{var}\big(S'_{n,j}(x)\big) \to f_{X_i}(x)\int u^{2j}K^{2(j+1)}(u)du,$$

as $n \to \infty$. Note that the integral is decreasing for increasing integers $j \geq 0$. Using the same line of argument to prove Lemma 1.3.6 it follows that

$$\mathbb{E}\,S'_{n,j}(x) = \begin{cases} \nu_{j+1}h_n f'_{X_i}(x) + \text{o}(h_n^2), & \text{for odd } j > 0; \\ \nu'_j f_{X_i}(x) + \mathcal{O}(h_n^2), & \text{for even, } j \geq 0, \end{cases}$$

with $\nu_j = \int u^j K^j(u)du$ and $\nu'_j = \int u^j K^{j+1}(u)du$. Define the random variables $Z'_{n,j,i}(x) = (X_i-x)^j K_{h_n}^{j+1}(X_i-x) - \mathbb{E}(X_1-x)^j K_{h_n}^{j+1}(X_1-x)$, with close resemblance to (A.12), and $V'_{n,j,k}(x)$ similar to (A.13). Follow the steps of the proof of Lemma 1.3.6 with only minor adjustments, then, at every continuity point x of $f_{X_i}(\cdot)$,

$$S'_{n,j}(x) = \begin{cases} \nu_{j+1}h_n f'_{X_i}(x) + \mathcal{O}\Big(\sqrt{\ln(n)/(nh_n)}\Big) & \text{almost surely, for odd } j > 0; \\ \nu'_j f_{X_i}(x) + \mathcal{O}(h_n^2) + \mathcal{O}\Big(\sqrt{\ln(n)/(nh_n)}\Big) & \text{almost surely, for even } j \geq 0. \end{cases}$$

Combining the above result with (A.29), $\lambda_n(x) = \mathcal{O}(h_n) + \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ almost surely, and the fact that $\nu'_0 = \int K(u)du = 1$, then at every continuity point x of $f_{X_i}(\cdot)$,

$$J_3(x) = f_{X_i}(x) + \mathcal{O}(h_n^2) + \mathcal{O}\Big(\sqrt{\ln(n)/(nh_n)}\Big) \qquad \text{almost surely,}$$

which completes the proof. □

A.5 Proof of Proposition 1.3.9

The proof is similar to the proof of the previous proposition. Note that, given the linearity constraint in (1.7), it follows that

$$\begin{aligned} J_2(x) &= \sum_{i=1}^{n}(m(X_i)-m(x))p_i(x;\lambda_n)K_{h_n}(X_i-x) \\ &= \sum_{i=1}^{n}\Big(m'(x)(X_i-x)+\frac{1}{2}m''(x)(X_i-x)^2+\mathrm{o}(h_n^2)\Big)p_i(x;\lambda_n)K_{h_n}(X_i-x) \\ &= \frac{m''(x)}{2}\sum_{i=1}^{n}(X_i-x)^2p_i(x;\lambda_n)K_{h_n}(X_i-x)+\mathrm{o}(h_n^2)\sum_{i=1}^{n}p_i(x;\lambda_n)K_{h_n}(X_i-x) \\ &= \frac{h_n^2m''(x)}{2}\big(S''_{n,0}(x)-\lambda_n(x)S''_{n,1}(x)+\lambda_n^2(x)S''_{n,2}(x)-...\big)+\mathrm{o}(h_n^2)J_3(x). \qquad \text{(A.30)} \end{aligned}$$

The second line follows from the approximation of $m(X_i)$ given in (1.2), Assumption 1.3.1.c), and the bounded support of the kernel function. The fourth line is due to the binomial representation of the probabilities $p_i(x;\lambda_n)$, the definition of $J_3(x)$ in (1.17), and $S''_{n,j}(x) = (nh_n)^{-1}\sum_{i=1}^{n}((X_i-x)/h_n)^{j+2}K^{j+1}((X_i-x)/h_n)$ for integers $j \geq 0$. Following similar arguments as in the proof of Lemmas 1.3.5 and 1.3.6 lead to

$$nh_n\mathrm{var}(S''_{n,j}(x)) \to f_{X_i}(x)\int u^{2(j+2)}K^{2(j+1)}(u)du,$$

as $n \to \infty$, and

$$S''_{n,j}(x) = \begin{cases} \mu'_{j+1}h_nf'_{X_i}(x)+\mathcal{O}\Big(\sqrt{\ln(n)/(nh_n)}\Big) & \text{almost surely, for odd } j>0; \\ \mu_{j+1}f_{X_i}(x)+\mathcal{O}(h_n^2)+\mathcal{O}\Big(\sqrt{\ln(n)/(nh_n)}\Big) & \text{almost surely, for even } j\geq 0, \end{cases}$$

at every continuity point x of $f_{X_i}(\cdot)$, with $\mu'_j = \int u^{j+2}K^j(u)du$ and $\mu_j = \int u^{j+1}K^j(u)du$, respectively. Combining this with (A.30), Theorem 1.3.7, and Proposition 1.3.8 it follows that at every continuity point x of $f_{X_i}(\cdot)$,

$$J_2(x) = \frac{h_n^2\mu_1 f_{X_i}(x)m''(x)}{2}\left(1+\mathcal{O}(h_n^2)+\mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right)\right) \qquad \text{almost surely,}$$

which completes the proof.

□

A.6 Proof of Lemma 1.3.10

I prove the statement by proving $|\phi(X_{i+1})| \leq \tau_n$ almost surely for $i \leq n$ and n sufficiently large using the Borel-Cantelli lemma. Note that

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(|\phi(X_{n+1})| > \tau_n) &\leq \sum_{n=1}^{\infty} \frac{\mathbb{E}\,|\phi(X_{n+1})|^s}{\tau_n^s} \\ &= \mathbb{E}\,|\phi(X_2)|^s \sum_{n=1}^{\infty} \tau_n^{-s} < \infty.\end{aligned}$$

The first line follows from Markov's inequality and the second from the finiteness of $\mathbb{E}\,|\phi(X_2)|^s$ (Assumption 1.3.1.b)) , stationarity, and the summability of τ_n^{-s}, with τ_n given in (1.22). Summability of τ_n^{-s} is established using the integral test, i.e.,

$$\int_n^{\infty} \frac{1}{y(\ln(\ln(y)))^2 \ln(y)} dy = -\frac{1}{\ln(\ln(y))}\bigg|_n^{\infty} = \frac{1}{\ln(\ln(n))} < \infty,$$

for $n > e$. Thus, $|\phi(X_{n+1})| \leq \tau_n$ with probability one for sufficiently large n. Since τ_n is increasing in n, $|\phi(X_{i+1})| \leq \tau_n$ with probability one for $i \leq n$ and n sufficiently large. Thus, the difference $R_{n,j}(x)$, given in (1.23), is o(1) with probability one.

□

A.7 Proof of Lemma 1.3.11

The proof bares some resemblance to the proof of Lemma 1.3.5. Define the random variable $U_{j,i}(x) = (\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})| \leq \tau_n\}} - m^{(t)}(X_i))(X_i - x)^j K_{h_n}^{j+1}(X_i - x)$ for integers $j \geq 0$. Then,

$$\begin{aligned}\operatorname{var}\Big(T_{n,j}^{(t)}(x)\Big) &= \frac{1}{n^2}\operatorname{var}\left(\sum_{i=1}^{n}\Big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})| \leq \tau_n\}} - m^{(t)}(X_i))(X_i - x)^j K_{h_n}^{j+1}(X_i - x)\Big)\right) \\ &= \frac{1}{n^2}\operatorname{var}\left(\sum_{i=1}^{n} U_{j,i}(x)\right) \\ &= \frac{1}{n}\operatorname{var}(U_{j,1}(x)) + \frac{2}{n}\sum_{i=2}^{n}\left(1 - \frac{i-1}{n}\right)\operatorname{cov}(U_{j,1}(x), U_{j,i}(x)) \\ &=: \Sigma_{j,1}(x) + \Sigma_{j,2}(x), \end{aligned} \tag{A.31}$$

where the third line follows from stationarity. To determine $nh_n\Sigma_{j,1}(x)$, with $\Sigma_{j,1}(x)$ given in (A.31), define the function $k_j(u) = u^j K^{j+1}(u)$, then, since $\mathbb{E}\, U_{j,i}(x) = 0$,

$$\begin{aligned}
nh_n\Sigma_{j,1}(x) &= \frac{1}{h_n}\,\mathbb{E}\bigg(\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\le\tau_n\}} - m^{(t)}(X_1)\Big)\bigg(\frac{X_1-x}{h_n}\bigg)^j K^{j+1}\bigg(\frac{X_1-x}{h_n}\bigg)\bigg)^2 \\
&= \frac{1}{h_n}\,\mathbb{E}\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\le\tau_n\}} - m^{(t)}(X_1)\Big)^2 k_j^2((X_1-x)/h_n)) \\
&= \frac{1}{h_n}\,\mathbb{E}\bigg(\mathbb{E}\Big(\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\le\tau_n\}} - m^{(t)}(X_1)\Big)^2\Big|X_1\Big)k_j^2((X_1-x)/h_n))\bigg) \\
&= \frac{1}{h_n}\int \mathbb{E}\Big(\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\le\tau_n\}} - m^{(t)}(z)\Big)^2\Big|X_1=z\Big) \\
&\qquad\qquad \times k_j^2((z-x)/h_n))f_{X_i}(z)dz \\
&= \frac{1}{h_n}\int \mathbb{E}\Big(\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\le\tau_n\}} - m^{(t)}(x+v)\Big)^2\Big|X_1=x+v\Big) \\
&\qquad\qquad \times k_j^2(v/h_n)f_{X_i}(x+v)dv \\
&\to \mathbb{E}\Big(\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\le\tau_n\}} - m^{(t)}(x)\Big)^2\Big|X_1=x\Big)f_{X_i}(x)\int k_j^2(u)du \\
&=: C(x)f_{X_i}(x)\int k_j^2(u)du, \qquad\qquad (A.32)
\end{aligned}$$

at every continuity point x of $C(\cdot)$ and $f_{X_i}(\cdot)$, as $n\to\infty$. The fifth equation follows from a change of variable, i.e., $v = z - x$. The second to last line is an application of Bochner's lemma given in Lemma B.5 in Appendix B and the continuity of the conditional variance (established in Lemma B.10). To apply Bochner's lemma define the function $H(u) = k_j^2(u)$, then $H(\cdot)$ fulfills the conditions i)–iii) of the lemma since the kernel function is bounded and has bounded support. Furthermore, define

$$g(x+v) = \mathbb{E}\bigg(\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\le\tau_n\}} - m^{(t)}(x+v)\Big)^2\Big|X_1=x+v\bigg)f_{X_i}(x+v).$$

Given the bounded support of the kernel function it suffices to check the finiteness of $\int|g(u)|du$ on the domain $\mathcal{D}(x) = \{r\in\mathbb{R} : x-h_n\le r\le x+h_n\}$ in order to apply the lemma. Thus,

$$\begin{aligned}
\int_{\mathcal{D}(x)}|g(u)|du &= \int_{\mathcal{D}(x)}\bigg|\,\mathbb{E}\bigg(\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\le\tau_n\}} - m^{(t)}(u)\Big)^2\Big|X_1=u\bigg)f_{X_i}(u)\bigg|du \\
&\le \int_{\mathcal{D}(x)}\mathbb{E}\bigg(\Big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\le\tau_n\}}\big| + \big|m^{(t)}(u)\big|\Big)^2\Big|X_1=u\bigg)f_{X_i}(u)du \\
&\le \sup_{u\in\mathcal{D}(x)}\Big\{\mathbb{E}\Big(\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\le\tau_n\}}\big| + C_4\big)^2\Big|X_1=u\Big)\Big\}\int_{\mathcal{D}(x)}f_{X_i}(u)du \\
&< \infty.
\end{aligned}$$

The second to last line follows from $\sup_{X_i\in\mathcal{D}(x)}|m(X_i)| \leq C_4$ (see Lemma B.4) and $|m^{(t)}(X_i)| \leq |m(X_i)|$ as well as $\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big| \leq \big|\phi(X_2)\big|$ and Assumption 1.3.1.b). Note that since $\mathbb{E}\,|\phi(X_{i+1})|^s < \infty$, for some $s > 2$ (Assumption 1.3.1.b)), it follows that $\sup_{u\in\mathcal{D}(x)}\mathbb{E}(|\phi(X_2)|^2|X_1 = u) < \infty$. Furthermore, define

$$g_n(x) = \frac{1}{h_n}\int H\bigg(\frac{v}{h_n}\bigg)g(x+v)dv,$$

then by virtue of Bochner's lemma the result in (A.32) follows.

To bound $nh_n\Sigma_{j,2}(x)$, with $\Sigma_{j,2}(x)$ defined in (A.31), note that

$$\begin{aligned}
nh_n\Sigma_{j,2}(x) &\leq 2h_n\sum_{i=2}^{n}\operatorname{cov}(U_{j,1}(x),U_{j,i}(x)) \\
&= 2h_n\sum_{i=2}^{\lfloor d_n\rfloor}\big|\operatorname{cov}(U_{j,1}(x),U_{j,i}(x))\big| \\
&\qquad + 2h_n\sum_{i=\lfloor d_n\rfloor+1}^{n}\big|\operatorname{cov}(U_{j,1}(x),U_{j,i}(x))\big| \\
&=: \Sigma_{j,21}(x) + \Sigma_{j,22}(x), \qquad\qquad (A.33)
\end{aligned}$$

where d_n is similar defined as in the proof of Lemma 1.3.5, fulfilling $d_n \to \infty$, $d_nh_n \to 0$, as $n \to \infty$. For each summand of $\Sigma_{j,21}(x)$ it follows that

$$\begin{aligned}
&h_n\big|\operatorname{cov}(U_{j,1}(x),U_{j,i}(x))\big| \\
&= \frac{1}{h_n}\bigg|\mathbb{E}\Big(\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}} - m^{(t)}(X_1)\Big) \\
&\qquad\qquad \times\Big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}} - m^{(t)}(X_i)\Big)k_j\bigg(\frac{X_1-x}{h_n}\bigg)k_j\bigg(\frac{X_i-x}{h_n}\bigg)\Big)\bigg| \\
&\leq \frac{1}{h_n}\,\mathbb{E}\,\bigg(\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big| + C_4\big) \\
&\qquad\qquad \times\big(\big|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\big| + C_4\big)\bigg|k_j\bigg(\frac{X_1-x}{h_n}\bigg)\bigg|\bigg|k_j\bigg(\frac{X_i-x}{h_n}\bigg)\bigg|\bigg) \\
&= \frac{1}{h_n}\,\mathbb{E}\,\bigg(\bigg|k_j\bigg(\frac{X_1-x}{h_n}\bigg)\bigg|\bigg|k_j\bigg(\frac{X_i-x}{h_n}\bigg)\bigg| \\
&\qquad\qquad \times\mathbb{E}\big(\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big| + C_4\big)\big(\big|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\big| + C_4\big)\big|X_1,X_i\big)\bigg) \\
&= \frac{1}{h_n}\int\bigg|k_j\bigg(\frac{z-x}{h_n}\bigg)\bigg|\bigg|k_j\bigg(\frac{w-x}{h_n}\bigg)\bigg|f_{X_1,X_i}(z,w)dzdw \\
&\qquad\times\mathbb{E}\big(\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big| + C_4\big)\big(\big|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\big| + C_4\big)\big|X_1 = z, X_i = w\big)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{h_n}\int \left|k_j\Big(\frac{z-x}{h_n}\Big)\right|\left|k_j\Big(\frac{w-x}{h_n}\Big)\right| f_{X_1,X_i}(z,w)dzdw \\
&\qquad \times \sup_{z,w\in\mathcal{D}(x)} \Big\{ \mathbb{E}\left(\left(\left|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\right| + C_4\right)\right. \\
&\qquad\qquad \left.\times \left(\left|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\right| + C_4\right)\Big| X_1 = z, X_i = w\right)\Big\} \\
&= h_n \int \left|k_j(u)\right|\left|k_j(v)\right| f_{X_1,X_i}(x+uh_n, x+vh_n)dudv \\
&\qquad \times \sup_{z,w\in\mathcal{D}(x)} \Big\{ \mathbb{E}\left(\left(\left|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\right| + C_4\right)\right. \\
&\qquad\qquad \left.\times \left(\left|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\right| + C_4\right)\Big| X_1 = z, X_i = w\right)\Big\} \\
&\leq C_5 h_n \left(\int \left|k_j(u)\right| du\right)^2 \sup_{z,w\in\mathcal{D}(x)} \Big\{ \mathbb{E}\left(\left(\left|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\right| + C_4\right)\right. \\
&\qquad\qquad \left.\times \left(\left|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\right| + C_4\right)\Big| X_1 = z, X_i = w\right)\Big\} \\
&\leq C_5 h_n \left(\int \left|k_j(u)\right| du\right)^2 \sup_{z,w\in\mathcal{D}(x)} \Big\{ \sqrt{\mathbb{E}\Big(\left(\left|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\right| + C_4\right)^2 \Big| X_1 = z\Big)} \\
&\qquad\qquad \times \sqrt{\mathbb{E}\Big(\left(\left|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\right| + C_4\right)^2 \Big| X_i = w\Big)}\Big\} \\
&= \overline{C} h_n,
\end{aligned}
$$

for integers $j \geq 0$ and suitable constant $\overline{C}$. For the result use $\sup_{X_i\in\mathcal{D}(x)} |m(X_i)| \leq C_4$ and $|m^{(t)}(X_i)| \leq |m(X_i)|$, changes in variables, i.e., $u = (z-x)/h_n$ and $v = (w-x)/h_n$, and Assumption 1.3.3.b). The last inequality follows from the Cauchy-Schwarz inequality. The last line is due to Assumption 1.3.1.b), the fact that $\left|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\right| \leq \left|\phi(X_{i+1})\right|$, and $|k_j(u)| \geq |k_{j+1}(u)|$. Thus it follows that

$$
\Sigma_{j,21}(x) \leq \overline{C} h_n \lfloor d_n \rfloor \leq \overline{C} h_n d_n \to 0, \tag{A.34}
$$

as $n \to \infty$ for integers $j \geq 0$.

To determine $\Sigma_{j,22}(x)$, defined in (A.33), use again Davydov's lemma, given in Lemma B.6, by first checking the following requirement

$$
\Big(\mathbb{E}\left|U_{j,1}(x)\right|^\delta\Big)^{1/\delta} = \frac{1}{h_n}\left(\mathbb{E}\left|\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}} - m^{(t)}(X_1)\Big)k_j\Big(\frac{X_1-x}{h_n}\Big)\right|^\delta\right)^{1/\delta}
$$

$$
\begin{aligned}
&\leq \frac{1}{h_n}\left(\int \left|k_j\left(\frac{z-x}{h_n}\right)\right|^{\delta} f_{X_i}(z)dz\right)^{1/\delta} \\
&\qquad\qquad \times \left(\sup_{z\in\mathcal{D}(x)}\left\{\mathbb{E}\Big(\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big|+C_4\big)^{\delta}\big|X_i=z\Big)\right\}\right)^{1/\delta} \\
&\leq \frac{1}{h_n^{1-1/\delta}}\left(\int |k_j(u)|^{\delta} f_{X_i}(x+uh_n)du\right)^{1/\delta} \\
&\qquad\qquad \times \left(\sup_{z\in\mathcal{D}(x)}\left\{\mathbb{E}\Big(\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big|+C_4\big)^{\delta}\big|X_i=z\Big)\right\}\right)^{1/\delta} \\
&= \frac{\overline{C}}{h_n^{1-1/\delta}},
\end{aligned}
$$

for integers $j \geq 0$ and with suitable constant $\overline{C}$. The arguments are similar to the ones used to derive $\Sigma_{j,21}(x)$. By assumption $s \geq \delta$ implying a finite supremum in the second to last line. It follows that $\mathbb{E}\,|U_{j,1}(x)|^{\delta} < \infty$ and by virtue of Davydov's lemma

$$
\Sigma_{j,22}(x) \leq \frac{\overline{C}}{h_n^{1-2/\delta}} \sum_{k=\lfloor d_n \rfloor}^{\infty} (\alpha(k))^{1-2/\delta} \leq \overline{C} \sum_{k=\lfloor d_n \rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta} \to 0, \tag{A.35}
$$

as $n \to \infty$. By adding (A.34) and (A.35) it follows that

$$
nh_n\Sigma_{j,2}(x) \leq \overline{C}h_n d_n + \overline{C} \sum_{k=\lfloor d_n \rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta} \to 0. \tag{A.36}
$$

Finally, combining (A.31), (A.32), and (A.36) proves the statement.

□

A.8 Proof of Lemma 1.3.12

Due to the similarity of the proof of Lemma 1.3.6 only the necessary steps are presented. Similar to $Z_{n,j,i}(x)$, given in (A.12), define

$$U_{j,i}(x) = \frac{1}{h_n}\Big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}} - m^{(t)}(X_i)\Big)\left(\frac{X_i - x}{h_n}\right)^j K^{j+1}\left(\frac{X_i - x}{h_n}\right),$$

for integers $j \geq 0$. Similar to (A.13) partition $\{1, ..., n\}$ into $2q_n$ consecutive blocks with each block containing σ_n elements and define

$$V_{n,j,k}(x) = \frac{1}{n}\sum_{i=(k-1)\sigma_n+1}^{k\sigma_n} U_{j,i}(x).$$

Note that σ_n is not necessarily equal to s_n, where s_n is used in (A.13). It follows that

$$\begin{aligned} T_{n,j}^{(t)}(x) =& \frac{1}{n}\sum_{i=1}^{n} U_{j,i}(x) \\ &= \sum_{k=1}^{q_n} V_{n,j,2k-1}(x) + \sum_{k=1}^{q_n} V_{n,j,2k}(x) + \frac{1}{n}\sum_{i=2q_n\sigma_n+1}^{n} U_{j,i}(x) \\ &=: W'_{n,j}(x) + W''_{n,j}(x) + W'''_{n,j}(x). \end{aligned}$$

Proving summability of

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} U_{j,i}(x)\right| > \varepsilon\right) &= \mathbb{P}\big(\big|W'_{n,j}(x) + W''_{n,j}(x) + W'''_{n,j}(x)\big| > \varepsilon\big) \\ &\leq \mathbb{P}\Big(\big|W'_{n,j}(x)\big| > \frac{\varepsilon}{3}\Big) + \mathbb{P}\Big(\big|W''_{n,j}(x)\big| > \frac{\varepsilon}{3}\Big) \\ &\quad + \mathbb{P}\Big(\big|W'''_{n,j}(x)\big| > \frac{\varepsilon}{3}\Big), \end{aligned} \tag{A.37}$$

for every $\varepsilon > 0$ establishes the statement of the lemma. Without specifying the details apply Bradley's coupling theorem to determine $\mathbb{P}(\big|W'_{n,j}(x)\big| > \frac{\varepsilon}{3})$, i.e.,

$$\begin{aligned} \mathbb{P}\Big(\big|W'_{n,j}(x)\big| > \frac{\varepsilon}{3}\Big) \leq& \mathbb{P}\left(\left|\sum_{k=1}^{q_n} V^*_{n,j,2k-1}(x)\right| > \frac{\varepsilon}{6}\right) \\ &+ \mathbb{P}\left(\left|\sum_{k=1}^{q_n} V^*_{n,j,2k-1}(x) - V_{n,j,2k-1}(x)\right| > \frac{\varepsilon}{6}\right), \end{aligned} \tag{A.38}$$

similar to (A.15). For the first term of (A.38) apply Bernstein's inequality and determine the bound b for $V^*_{n,j,2k-1}(x)$ and $\sum_{k=1}^{q_n} \mathbb{E}(V^*_{n,j,2k-1}(x))^2$. For the first part note that

$$\begin{aligned}\left|U_{j,i}(x)\right| &= \frac{1}{h_n}\left|\left(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}} - m^{(t)}(X_i)\right)\left(\frac{X_i-x}{h_n}\right)^j K^{j+1}\left(\frac{X_i-x}{h_n}\right)\right| \\ &\leq \frac{2\tau_n}{h_n}\left|\frac{X_i-x}{h_n}\right|^j K^{j+1}\left(\frac{X_i-x}{h_n}\right) \\ &\leq \frac{\overline{C}\tau_n}{h_n},\end{aligned}$$

with suitable constant $\overline{C}$, implying

$$\left|V^*_{n,j,2k-1}(x)\right| \leq \frac{\overline{C}\sigma_n\tau_n}{nh_n} =: b,$$

because $V_{n,j,2k-1}(x)$ and $V^*_{n,j,2k-1}(x)$ have the same distribution. Note that τ_n appears in the numerator, thus b is different to (A.16). For the second part, i.e., to bound $\sum_{k=1}^{q_n} \mathbb{E}(V^*_{n,j,2k-1}(x))^2$ similar derivations that led (A.17) and Lemma 1.3.11 lead to

$$\sum_{k=1}^{q_n} \mathbb{E}(V^*_{n,j,2k-1}(x))^2 \leq \frac{1}{nh_n} C(x) f_{X_i}(x) \int u^{2j} K^{2(j+1)}(u)du(1+\mathrm{o}(1)).$$

Choosing $\varepsilon = \varepsilon_n = C_\varepsilon\sqrt{\ln(n)/(nh_n)}$ and $\sigma_n = \lfloor \tau_n^{-1}\sqrt{nh_n/\ln(n)}\rfloor$ it follows that

$$\mathbb{P}\left(\left|\sum_{k=1}^{q_n} V^*_{n,j,2k-1}(x)\right| > \frac{\varepsilon_n}{6}\right) \leq 2n^{-\beta}, \tag{A.39}$$

with $\beta = C_\varepsilon^2/(\overline{C}(x)+\overline{C})$ and $\overline{C}(x)$ absorbing the integral term of (1.24), $C(x)$, and all other constants. By choosing C_ε sufficiently large (A.39) is summable.

For the second term of (A.38) similar derivations that led to (A.21) result in

$$\begin{aligned}\sum_{k=1}^{q_n} \mathbb{P}\left(\left|V^*_{n,j,2k-1}(x) - V_{n,j,2k-1}(x)\right| > \frac{\varepsilon_n}{6q_n}\right) &\leq \overline{C}q_n\sqrt{\max\left\{\frac{6q_n\|V_{n,j,2k-1}(x)\|_\infty}{\varepsilon_n}, 1\right\}}\alpha(\sigma_n+1) \\ &\leq \overline{C}q_n\sqrt{\max\left\{\frac{\overline{C}q_n\sigma_n\tau_n}{nh_n\varepsilon_n}, 1\right\}}\alpha(\sigma_n+1) \\ &\leq \frac{n\overline{C}}{\sigma_n}\sqrt{\frac{\tau_n}{h_n\varepsilon_n}}\alpha(\sigma_n+1) \\ &\leq \overline{C}\,\tau_n^{3/2}\left(\ln(n)\left(\frac{n}{h_n}\right)^3\right)^{1/4}\alpha(\sigma_n) \\ &= \overline{C}\,\Xi_n,\end{aligned} \tag{A.40}$$

where Ξ_n is summable by assumption. Because the results for the remaining two terms in (A.37) are similar it is easy to see, using (A.37), (A.38), (A.39) and (A.40), that

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} U_{j,i}(x)\right| > \varepsilon_n\right) \leq 6n^{-\beta} + \overline{C}\,\Xi_n,$$

is summable by choosing C_ε appropriately large. By virtue of the Borel-Cantelli lemma the statement

$$T_{n,j}^{(t)}(x) = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \text{almost surely,}$$

follows, therefore proving the lemma.

□

A.9 Proof of Proposition 1.3.13

Since $J_1(x) = \sum_{i=1}^{n}(\phi(X_{i+1}) - m(X_i))p_i(x;\lambda_n)K_{h_n}(X_i - x)$ it is easy to see that

$$\begin{aligned} J_1(x) &= T_{n,0}(x) - \lambda_n(x)T_{n,1}(x) + \lambda^2(x)T_{n,2}(x) - \lambda^3(x)T_{n,3}(x) + ... \\ &= T_{n,0}^{(t)}(x) - \lambda_n(x)T_{n,1}^{(t)}(x) + \lambda^2(x)T_{n,2}^{(t)}(x) - \lambda^3(x)T_{n,3}^{(t)}(x) + ..., \end{aligned}$$

with $T_{n,j}(x)$ and $T_{n,j}^{(t)}(x)$ defined in (1.20) and (1.21), respectively. The first equation follows from the binomial representation of the probabilities $p_i(x;\lambda_n)$. The second equation follows from Lemma 1.3.10. Since $\lambda_n(x) = \mathcal{O}(h_n) + \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ with probability one, due to Theorem 1.3.7, and $T_{n,j}^{(t)}(x) = \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ with probability one for integers $j \geq 0$, due to Lemma 1.3.12, it follows that $J_1(x) = \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ almost surely.

□

A.10 Proof of Theorem 1.3.14

Since $J_3(x) = f_{X_i}(x)(1 + \mathcal{O}(h_n^2) + \mathcal{O}(\sqrt{\ln(n)/(nh_n)}))$ almost surely (Proposition 1.3.8), it follows that at every continuity point x of $f_{X_i}(\cdot)$,

$$\frac{1}{J_3(x)} = \frac{1}{f_{X_i}(x)} \left(1 + \mathcal{O}(h_n^2) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \right) \quad \text{almost surely.}$$

The above expression is well-defined due to Assumption 1.3.3.a). Using this together with $J_2(x) = \mathcal{O}(h_n^2)$ almost surely (Proposition 1.3.9), it follows that

$$\frac{J_2(x)}{J_3(x)} = \mathcal{O}(h_n^2) \quad \text{almost surely.} \tag{A.41}$$

Since $J_1(x) = \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ with probability one (Proposition 1.3.9), similar arguments show that

$$\frac{J_1(x)}{J_3(x)} = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \quad \text{almost surely.} \tag{A.42}$$

By substituting (A.41) and (A.42) into (1.16) the statement of the theorem, i.e.,

$$\widehat{m}(x) - m(x) = \mathcal{O}(h_n^2) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \quad \text{almost surely,}$$

follows.

□

A.11 Proof of Corollary 1.3.15

Note that $h_n \sim (\ln(n)/n)^{1/5}$ is the optimal bandwidth, i.e., the bandwidth for which the convergence of $\widehat{m}(x) - m(x)$ is fastest. This can be easily seen by a simple maximization of $h_n^2 + \sqrt{\ln(n)/(nh_n)}$ with respect to the bandwidth. Inserting the optimal bandwidth into the expression of Theorem 1.3.14 proves the statement.

□

B Additional lemmas

Lemma B.1. *Given Assumption 1.3.2 it follows that* $C_2 = \sup_{u\in\mathbb{R}} uK(u) < 1$.

Proof. I prove the statement by contradiction and a simple geometrical argument. Suppose there exists one $u_1 \in supp(K)$ such that $u_1K(u_1) \geq 1$, then $u_1 \notin \{[-1,0]\} \cup \{1\}$ because for these value $u_1K(u_1) \leq 0$. Let $u_1 \in supp(K) \setminus (\{[-1,0]\} \cup \{1\})$ such that $u_1K(u_1) \geq 1$, then, given the symmetry of $K(\cdot)$, the area of the rectangle $2u_1K(u_1)$ is at least 2. But since $K(\cdot)$ is unimodal this implies that $\int_{-1}^{1} K(u)du > 2$ which contradicts the assumption of $K(\cdot)$ being a density.

□

Lemma B.2. *Assume that Assumption 1.3.2 holds, then* $|u|K(u) < \infty$ *for* $u \in \mathbb{R}$.

Proof. The proof is simple. Since the kernel function is symmetric and has bounded support it follows that $|u| \leq 1$ and $K(u) \leq K(0)$ for $u \in \mathbb{R}$. Thus, $|u|K(u) \leq K(0) < \infty$.

□

Lemma B.3. *Assume that Assumption 1.3.2 holds, then* $\int |u|K(u) < \infty$ *for* $u \in \mathbb{R}$.

Proof. Due to the proof of Lemma B.2 it follows that $\int |u|K(u)du \leq \int K(u)du = 1$ because the kernel function is a density.

□

Lemma B.4. *Assume that Assumptions 1.3.1.c) and 1.3.2 hold, then* $|m(X_i)|$ *is bounded in the neighborhood* $X_i \in \mathcal{D}(x)$, *with* $\mathcal{D}(x) = \{r \in \mathbb{R} : x - h_n \leq r \leq x + h_n\}$.

Proof. Due to the bounded support of the kernel function it suffices to consider $m(X_i)$ only for $X_i \in \mathcal{D}(x)$. Then, according to (1.2) with $q = 1$,

$$\begin{aligned}
\sup_{X_i\in\mathcal{D}(x)} |m(X_i)| &= \sup_{X_i\in\mathcal{D}(x)} \left| m(x) + m'(x)(X_i - x) + \frac{1}{2}m''(x)(X_i - x)^2 + \mathrm{o}(h_n^2) \right| \\
&\leq \big|m(x)\big| + \big|m'(x)\big| \sup_{X_i\in\mathcal{D}(x)} \big\{\big|X_i - x\big|\big\} \\
&\qquad + \frac{1}{2}\big|m''(x)\big| \sup_{X_i\in\mathcal{D}(x)} \big\{(X_i - x)^2\big\} + \mathrm{o}(h_n^2) \\
&\leq \big|m(x)\big| + h_n\big|m'(x)\big| + \frac{h_n^2}{2}\big|m''(x)\big| + \mathrm{o}(h_n^2) \\
&< \infty.
\end{aligned}$$

The last line follows from Assumption 1.3.1.c).

□

Lemma B.5 (Bochner's lemma). *Suppose $H(\cdot)$ is a Borel measurable function satisfying the conditions*

i) $\sup_{v\in\mathbb{R}} |H(v)| < \infty$,

ii) $\int |H(v)|dv < \infty$,

iii) $|vH(v)| \to 0$ *as* $v \to \infty$.

Let $g(v)$ satisfy $\int |g(v)|dv < \infty$ and $h_n \to 0$ as $n \to \infty$. Let

$$g_n(x) = \frac{1}{h_n} \int H\left(\frac{v}{h_n}\right) g(x+v)dv,$$

then at every continuity point x of $g(\cdot)$,

$$g_n(x) \to g(x) \int H(v)dv,$$

as $n \to \infty$.

Proof. See Parzen (1962, pp. 1067–1068). Bosq and Blanke (2007, Chaper 6) provide a series of applications.

□

Lemma B.6 (Davydov's lemma). *Let X and Y be two random variables defined a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let α be the strong mixing coefficient measuring the dependence of X and Y. If $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|Y|^q < \infty$ for $p, q \geq 1$ and $p^{-1} + q^{-1} < 1$, it follows that*

$$\left|cov(X,Y)\right| \leq 8\alpha^{1-1/p-1/q} \|X\|_p \|Y\|_q,$$

with $\|X\|_p = \left(\mathbb{E}\left|X\right|^p\right)^{1/p}$ and $\|Y\|_q = \left(\mathbb{E}\left|Y\right|^q\right)^{1/q}$.

Remark B.7. The original statement of the lemma according to Davydov (1968) the constant is 12 instead of 8.

Proof. See Hall and Heyde (1980, p. 278).

□

Lemma B.8 (Bernstein inequality). *Let $\{X_i\}_{i=1}^n$ be independent zero-mean random variables. Suppose that $|X_i| \leq b$ almost surely, for all i. Then, for all positive ε,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| > \varepsilon\right) \leq 2\exp\left\{-\frac{\frac{1}{2}\varepsilon^2}{\sum_{i=1}^n \mathbb{E}\, X_i^2 + \frac{1}{3}b\varepsilon}\right\}.$$

Proof. See Pollard (1984, pp. 192–193).

□

Lemma B.9 (Bradley's coupling theorem). *Let (X, Y) be a $\mathbb{R}^d \times \mathbb{R}$-valued random vector, with integers $d > 0$, such that $0 < \varepsilon \leq \|Y\|_\infty < \infty$. Suppose U is a standard uniform random variable independent of (X, Y). Then there exists a real-valued random variable $Y^* = \psi(X, Y, U)$ where $\psi(\cdot)$ is a measurable function $\mathbb{R}^d \times \mathbb{R} \times [0, 1]$ into $\mathbb{R}$ such that*

i) Y^* is independent of X,

ii) Y^* and X have the same distribution,

iii)

$$\mathbb{P}\big(|Y^* - Y| > \varepsilon\big) \leq 18\sqrt{\frac{\|Y\|_\infty}{\varepsilon}} \sup_{\substack{A\in\sigma(X)\\ B\in\sigma(Y)}} \big|\,\mathbb{P}(A \cap B) - \mathbb{P}(A)\,\mathbb{P}(B)\big|.$$

Proof. See Bradley (1983, pp. 74–76) and the discussion after his Theorem 3. □

Lemma B.10. *Assume that Assumptions 1.3.1.c) and 1.3.3.c) hold, then, for fixed $\tau > 0$,*

$$var\big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau\}}\big|X_1 = u\big),$$

is continuous at $u = x$.

Proof. Note that

$$\begin{aligned}
&\operatorname{var}\big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau\}}\big|X_1 = u\big)\\
&\quad= \mathbb{E}\left(\big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau\}} - \mathbb{E}\big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau\}}\big|X_1 = u\big)\big)^2\big|X_1 = u\right)\\
&\quad= \mathbb{E}\left(\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau\}} - m^{(t)}(u)\Big)^2\Big|X_1 = u\right)\\
&\quad= \mathbb{E}\big(\phi^2(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau\}}\big|X_1 = u\big) - \big(\mathbb{E}\big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau\}}\big|X_1 = u\big)\big)^2.
\end{aligned}$$

By virtue of dominated convergence as well as Assumptions 1.3.3.c) and 1.3.1.d) it follows that

$$\begin{aligned}
\mathbb{E}\big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau\}}\big|X_1 = u\big) &= \mathbb{E}\big((m(X_1) + \sigma(X_1)\epsilon)\mathbb{1}_{\{|m(X_1)+\sigma(X_1)\epsilon|\leq\tau\}}\big|X_1 = u\big)\\
&= m(u)\,\mathbb{E}\big(\mathbb{1}_{\{|m(u)+\sigma(u)\epsilon|\leq\tau\}}\big)\\
&\to m(x)\,\mathbb{E}\big(\mathbb{1}_{\{|m(x)+\sigma(x)\epsilon|\leq\tau\}}\big)\\
&= \mathbb{E}\big((m(X_1) + \sigma(X_1)\epsilon)\mathbb{1}_{\{|m(X_1)+\sigma(X_1)\epsilon|\leq\tau\}}\big|X_1 = x\big)\\
&= \mathbb{E}\big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau\}}\big|X_1 = x\big),
\end{aligned}$$

as $u \to x$. Continuity of the first term follows by similar arguments. □

References

Bao, Y., T.-H. Lee, and B. Saltoğlu (2006): "Evaluating Predictive Performance of Value-at-Risk Models in Emerging Markets: A Reality Check," *Journal of Forecasting*, 25, 101–128.

Basrak, B., R. A. Davis, and T. Mikosch (2002): "Regular Variation of GARCH Processes," *Stochastic Processes and their Applications*, 99, 95–115.

Bosq, D. and D. Blanke (2007): *Inference and Prediction in Large Dimensions*, Chichester: John Wiley & Sons.

Bradley, R. C. (1983): "Approximation Theorems for Strongly Mixing Random Variables," *Michigan Mathematical Journal*, 24, 59–70.

——— (2005): "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions," *Probability Surveys*, 2, 107–144.

Cai, Z. (2001): "Weighted Nadaraya-Watson Regression Estimation," *Statistics & Probability Letters*, 51, 307–318.

——— (2002): "Regression Quantiles for Time Series," *Econometric Theory*, 18, 169–192.

Cai, Z. and X. Wang (2008): "Nonparametric Estimation of Conditional VaR and Expected Shortfall," *Journal of Econometrics*, 147, 120–130.

Chen, S. X. and P. Hall (1993): "Smoothed Empirical Likelihood Condfidence Intervals for Quantiles," *The Annals of Statistics*, 21, 1166–1181.

Cheng, P. E. (1995): "A Note on Strong Convergence Rates in Nonparametric Regression," *Statistics & Probability Letters*, 24, 357–364.

Chesney, M., G. Reshetar, and M. Karaman (2011): "The Impact of Terrorism on Financial Markets: An Empirical Study," *Journal of Banking & Finance*, 35, 253–267.

Cleveland, W. S. (1979): "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Davidson, J. (1994): *Stochastic Limit Theory*, Oxford: Oxford University Press.

Davydov, Y. A. (1968): "Convergence of Distributions Generated by Stationary Stochastic Processes," *Theory of Probability & Its Applications*, 13, 691–696.

Doukhan, P. (1994): *Mixing: Properties and Examples*, New York: Springer-Verlag.

Fan, J. (1993): "Local Linear Regression Smoothers and their Minimax Efficiencies," *The Annals of Statistics*, 21, 196–216.

Fan, J., T. Gasser, I. Gijbels, M. Brockmann, and J. Engel (1995): "On Nonparametric Estimation via Local Polynomial Regression," *Working paper University of Louvain.*

Fan, J. and I. Gijbels (1992): "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 29, 2008–2036.

——— (1996): *Local Polynomial Modelling and its Applications*, Boca Raton: Chapman & Hall/CRC.

Fryzlewicz, P. and S. Subba Rao (2011): "Mixing Properties of ARCH and Time-Varying ARCH Processes," *Bernoulli*, 17, 320–346.

Hall, P. and C. C. Heyde (1980): *Martingale Limit Theory and its Application*, New York: Academic Press.

Hall, P. and B. Presnell (1999): "Intentionally Biased Bootstrap Methods," *Journal of the Royal Statistical Society: Series B*, 61, 143–158.

Hall, P., R. C. L. Wolff, and Q. Yao (1999): "Methods for Estimating a Conditional Distribution Function," *Journal of the American Statistical Association*, 94, 154–163.

Hart, J. D. (1996): "Some Automated Methods of Smoothing Time-Dependent Data," *Nonparametric Statistics*, 6, 115–142.

Hastie, T. and C. Loader (1993): "Local Regression: Automatic Kernel Carpentry," *Statistical Science*, 8, 120–143.

Kato, K. (2012): "Weighted Nadaraya-Watson Estimation of Conditional Expected Shortfall," *Journal of Financial Econometrics*, 10, 265–291.

Kristensen, D. (2009): "Uniform Convergence Rates of Kernel Estimators with Heterogeneous Dependent Data," *Econometric Theory*, 25, 1433–1445.

Li, Q. and S. Racine (2006): *Nonparametric Econometrics. Theory and Practice*, Princeton: Princeton University Press.

Lindner, A. M. (2009): "Stationarity, Mixing, Distributional Properties and Moments of GARCH(p,q)-Processes," in *Handbook of Financial Time Series*, ed. by T. G. Andersen, R. D. Davis, , J.-P. Kreiß, and T. Mikosch, Berlin: Spinger-Verlag, 43–69.

Lu, Z. and P. Cheng (1997): "Distribution-free Strong Consistency for Nonparametric Kernel Regression Involving Nonlinear Time Series," *Journal of Statistical Planning and Inference*, 65, 67–86.

Mack, Y. P. and B. W. Silverman (1982): "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61, 405–415.

Masry, E. (1996a): "Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates," *Journal of Time Series Analysis*, 17, 571–599.

——— (1996b): "Multivariate Regression Estimation: Local Polynomial Fitting for Time Series," *Stochastic Processes and their Application*, 65, 81–101.

Nadaraya, E. A. (1964): "On Estimating Regression," *Theory of Probability and Its Applications*, 9, 141–142.

Pagan, A. and A. Ullah (1999): *Nonparametric Econometrics*, Cambridge: Cambridge University Press.

Parzen, E. (1962): "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, 33, 1065–1076.

Pollard, D. (1984): *Convergence of Stochastic Processes*, New York: Springer-Verlag.

Rosenblatt, M. (1956): "A Central Limit Theorem and a Strong Mixing Condition," *Proceedings of the National Academy of Sciences of the United States of America*, 42, 43–47.

Sarda, P. and P. Vieu (2000): "Kernel Regression," in *Smoothing and Regression: Approaches, Computation, and Application*, ed. by M. G. Schimek, New York: John Wiley & Sons, chap. 3, 43–70.

Steikert, K. U. (2014): "A Local Bootstrap Procedure to Select the Bandwidth for the Weighted Nadaraya-Watson Estimator in Case of Weakly Dependent Data," *Working paper University of Zurich.*

Stone, C. J. (1977): "Consistent Nonparametric Regression," *The Annals of Statistics*, 5, 595–645.

——— (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053.

Tay, A. S. and C. Ting (2008): "Intraday Stock Prices, Volume, and Duration: A Nonparametric Conditional Density Analysis," in *High frequency financial econometrics: recent developments*, ed. by L. Bauwens, W. Pohlmeier, and D. Veredas, Heidelberg: Physica-Verlag, 253–268.

Tran, L. T. (1990): "Kernel Density Estimation under Dependence," *Statistics & Probability Letters*, 10, 193–201.

Walk, H. (2010): "Strong Consistency of Kernel Estimates of Regression Function under Dependence," *Statistics & Probability Letters*, 80, 1147–1156.

Watson, G. S. (1964): "Smooth Regression Analysis," *Sankhyā*, 26, 359–372.

Yu, K. and M. C. Jones (1998): "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228–237.

Chapter 2

Uniform Strong Consistency and Rates of Convergence for the Weighted Nadaraya-Watson Estimator for Strongly Mixing Processes

2.1 Introduction

In this manuscript I establish uniform strong consistency on compact subsets of $\mathbb{R}$, along with rates of convergence, for the weighted Nadaraya-Watson estimator for functions of weakly dependent data. Uniform consistency, which is stronger than pointwise consistency, is an important property and needed because it permits further research regarding consistency of estimation methods in which the weighted Nadaraya-Watson estimator is embedded. Examples of such are two-stage, semiparametric, or bootstrap estimators, among others. Semiparametric estimators are useful because in practice, the curse of dimensionally can, in case of multivariable models, render unreliable nonparametric estimations if the number of data points is insufficient (see Fan and Yao (2005, pp. 314–317)). The following two examples provide reason why uniform consistency for estimators is needed.

For the first example let $\{X_i\}_{i=1}^n$ be a strictly stationary random sequence defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consider a two-stage estimator, which in the first stage depends on a nonparametric estimator of the conditional mean of X_{i+1} given X_i, denoted by $m(X_i)$. In the second stage, the estimator depends on the residual $\widehat{\epsilon}_i$ of the first stage. Proving consistency in this case is difficult because $\widehat{\epsilon}_i - \epsilon_i = m(X_i) - \widehat{m}(X_i)$ depends on the random variable X_i. Therefore pointwise consistency of $\widehat{m}(x)$, for $x \in \mathbb{R}$, does not suffice to claim $m(X_i) - \widehat{m}(X_i) \to 0$, as $n \to \infty$. A solution to the problem is to establish $\widehat{m}(x) \to m(x)$ uniformly on $\mathbb{S}$, where $\mathbb{S} = \{r \in \mathbb{R} : |r| \leq c\}$ and c satisfying $|X_i| \leq c$, as $n \to \infty$. Thus, if the estimator $\widehat{m}(x)$ is uniformly consistent on compact subsets of $\mathbb{R}$ proving consistency of the two-stage estimator follows.

A second example for which uniform consistency is important are bootstrap estimators. To prove consistency of the distribution of $T_n = T_n(X_1, X_2, ..., X_n)$, depending on the estimator $\widehat{m}(x)$, suppose that the underlying data are a random sequence with unknown joint cumulative distribution function (CDF) F_0. Let $F \in \mathcal{F}$ denote a generic element of the class of finite-dimensional and continuous CDFs $\mathcal{F}$ and denote $G_n(\cdot, F)$ the exact distribution of T_n when the underlying data are sampled from F. The object of interest is therefore $G_n(\cdot, F_0)$. Asymptotic methods replace this unknown distribution with the asymptotic distribution denoted by $G_\infty(\cdot, F_0)$. The bootstrap estimator on the other hand replaces the unknown exact CDF of the data, F_0, by a consistent estimator $\widehat{F}_n$. If F_0 is continuous on $\mathbb{R}$, then $\widehat{F}_n$ converges uniformly to F_0 on $\mathbb{R}$ due to the Glivenko-Cantelli theorem (see van der Vaart (1998, p. 266)). For the bootstrap estimator to make sense $G_n(\cdot, \widehat{F}_n)$

must be uniformly close to $G_\infty(\cdot, F_0)$ for large n. This is because the unknown $G_n(\cdot, F_0)$ is uniformly close to the asymptotic distribution $G_\infty(\cdot, F_0)$. Thus, uniform consistency results help in proving consistency of bootstrap estimators. For more details on bootstrap estimators see Efron and Tibshirani (1994). For consistency results of bootstrap estimators, in particular in combination with data dependence, see Lahiri (2003). The results of this manuscript provide part of the theoretical foundations necessary for the local bootstrap procedure to select the bandwidth presented in Steikert (2014a). The framework extends the setting of Paparoditis and Politis (2000), using the weighted Nadaraya-Watson estimator instead of the ordinary Nadaraya-Watson estimator, to estimate the conditional CDF to generate bootstrap samples. This extension is important because of the favorable bias properties of this particular estimator.

For the present manuscript I consider weakly dependent data; in particular, I assume strongly mixing random sequences. This assumption is widely adopted in the literature with the process of classification of stochastic processes being strongly mixing still continuing. Examples of strongly mixing processes are finite-dependent processes, types of ARMA processes (see Davidson (1994, pp. 219–228) for sufficient conditions), classes of Markov chains (Bradley (2005, pp. 117–122)), linear GARCH processes (see Basrak et al. (2002, Theorem 3.1)), as well as (non-) stationary ARCH processes (see Fryzlewicz and Subba Rao (2011)). The result represents the first uniform consistency results for the weighted Nadaraya-Watson estimator. Complementing the result I also provide the convergence rate which is optimal in the sense of Stone (1982) and thus it is the best rate possible to attain. An equivalent uniform rates are established in Masry (1996) for local polynomial estimators in a similar setting. Hansen (2008) establishes uniform weak and strong consistency on bounded and slowly expanding sets for a general class of kernel estimators. These include kernel density and local polynomial estimators, among others. The present manuscript also provides a detailed analysis in case of a polynomial strong mixing coefficient. This is of practical importance because, depending on the speed of convergence of the bandwidth, it facilitates the determination of a rate of decay for which the results hold. Although a time series setting is considered all the established results hold for a more general setting. I highlight the differences throughout the text as well as in the proofs if necessary.

The remainder of this manuscript is organized as follows. Section 2.2 introduces and discusses the weighted Nadaraya-Watson estimator. Uniform strong consistency of the estimator is established in Section 2.3, followed by concluding remarks in

Section 2.4. All proofs are given in Appendix A. Moreover, Appendix B provides supplemental lemmas.

2.2 The weighted Nadaraya-Watson estimator

Let $\{X_i\}_{i=1}^{n+1}$ be a strictly stationary real valued time series defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Moreover, let $\phi(\cdot)$ be an arbitrary Borel-measurable function fulfilling $\mathbb{E}\,|\phi(X_{i+1})| < \infty$. Define the regression function $m(x)$ as the expected value of $\phi(X_{i+1})$ given $X_i = x$[1], i.e.,

$$m(x) = \mathbb{E}(\phi(X_{i+1})|X_i = x). \tag{2.1}$$

Considering the response function $\phi(\cdot)$ allows the representation of a variety of statistics such as the one-step ahead prediction ($\phi(X_{i+1}) = X_{i+1}$), raw moments thereof ($\phi(X_{i+1}) = X_{i+1}^j$ for integers $j > 0$), as well as conditional probabilities of the former ($\phi(X_{i+1}) = \mathbb{1}_{(-\infty,y]}(X_{i+1})$ for some $y \in \mathbb{R}$).

Note that I focus on a time series framework. This setting is a special case of a more general setting. All the presented results hold for processes of the form $\{(X_i, Y_i)\}_{i=1}^n$ because by setting $Y_i = X_{i+1}$ the presented framework emerges. In the following section I therefore add the assumptions necessary to prove the results for the general case. In what follows and in particular in the proofs below I complement the text for this case if necessary.

By approximating the unknown stochastic function $m(X_i) = \mathbb{E}(X_{i+1}|X_i)$ by a polynomial of total order q leads to the estimation of (2.1). Instead of a global approximation, as in the parametric case, a local approximation at x for the nonparametric estimation is considered, given a prespecified size of the local neighborhood, called the bandwidth, h_n. Assuming the existence of the $(q+1)$-th derivative of (2.1) at x the approximation reads

$$m(X_i) \approx \sum_{k=0}^{q} \frac{1}{k!} m^{(k)}(x)(X_i - x)^k. \tag{2.2}$$

To estimate the local coefficients $m^{(k)}(x)/k!$, for $k = 0, 1, ..., q$, the polynomial is

[1]Extending the framework to allow for lagged variables or a vector thereof as a conditioning variable is straightforward.

fitted locally by a weighted polynomial regression, i.e., minimizing

$$\sum_{i=1}^{n}\left(\phi(X_{i+1})-\sum_{k=0}^{q}\frac{1}{k!}m^{(k)}(x)(X_i-x)^k\right)^2 p_i(x;\lambda_n)K_{h_n}(X_i-x), \tag{2.3}$$

with respect to $m^{(k)}(x)/k!$ leads to the various nonparametric estimators. The functions $K_{h_n}(x)=K(x/h_n)/h_n$, with $K(\cdot)$ being the kernel function, and $p_i(x;\lambda_n)$ are both nonnegative weight functions. The estimator resulting from solving the minimization problem in (2.3) for $q=0$[2], with optimally selected probabilities, defines the weighted Nadaraya-Watson estimator, which reads

$$\widehat{m}(x)=\frac{\sum_{i=1}^{n}p_i(x;\lambda_n)K_{h_n}(X_i-x)\phi(X_{i+1})}{\sum_{i=1}^{n}p_i(x;\lambda_n)K_{h_n}(X_i-x)}. \tag{2.4}$$

The probabilities, $p_i(x;\lambda_n)$, are optimally determined by maximizing the empirical log-likelihood, i.e., $\sum_{i=1}^{n}\ln(p_i(x;\lambda_n))$, subject to the following three constraints

$$p_i(x;\lambda_n)\geq 0 \quad \text{and} \quad \sum_{i=1}^{n}p_i(x;\lambda_n)=1, \tag{2.5}$$

as well as

$$\sum_{i=1}^{n}(X_i-x)p_i(x;\lambda_n)K_{h_n}(X_i-x)=0. \tag{2.6}$$

The two conditions in (2.5) ensure that $p_i(x;\lambda_n)$ are probabilities. The third condition, (2.6), invokes a favorable bias property inherent to the local linear estimator. The weighted Nadaraya-Watson estimator therefore reproduces the superior bias properties of local linear estimator while, in the case of estimating the conditional CDF, preserving the property that the ordinary Nadaraya-Watson estimator is always a distribution function. To determine the probabilities, $p_i(x;\lambda_n)$, let $\lambda_n(x)$ denote the Lagrange multiplier for condition (2.6) of the reduced optimization problem. By maximizing the empirical log-likelihood given the two constraints, the strict positivity constraint of the probabilities is implicitly imposed by the objective

[2]In fact, minimizing (2.3) for $q=1$ results in the same estimator $\widehat{m}(x)$ if the constraint (2.6) is applied when solving the normal equations.

function, the probabilities read

$$p_i(x;\lambda_n) = \frac{1}{n(1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x))}, \tag{2.7}$$

for which a closed form solution does not exist because of the underdetermined system of equations of first-order conditions of the Lagrange optimization. For a detailed derivation see Fan and Yao (2005, pp. 456–457) or Li and Racine (2006, pp. 186–189). Given (2.7), the optimization problem to determine the probabilities can be reduced leading to the minimization of $L_n(x;\lambda_n) = \sum_{i=1}^n \ln(1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x))$ with respect to $\lambda_n(x)$. The first-order condition of this problem reads

$$L_n'(x;\lambda_n) = \frac{1}{nh_n}\sum_{i=1}^n \frac{(X_i-x)K_{h_n}(X_i-x)}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)} = 0, \tag{2.8}$$

leading to a unique $\lambda_n(x)$ which may be determined by the Newton-Raphson algorithm. The multiplication by $(nh_n)^{-1}$ facilitates the proofs in the appendix.

Applications of the estimator are found, e.g., in Cai (2002) for quantile regression estimators. Kato (2012) uses the estimator to propose an estimator of the conditional expected shortfall. A similar problem, as well as the estimation of the conditional Value-at-Risk, is discussed in Cai and Wang (2008). Bao et al. (2006) evaluate the predictive performance of the estimator in Value-at-Risk models. Tay and Ting (2008) investigate the CDF of high-frequency price changes conditional on trading volume and duration between trades. Steikert (2014a) uses the estimator for a local bootstrap procedure to select the bandwidth for the weighted Nadaraya-Watson estimator.

2.3 Uniform strong consistency

The following subsection discusses the assumptions used in order to prove uniform strong consistency for the weighted Nadaraya-Watson estimator on compact subsets of $\mathbb{R}$. Section 2.3.2 presents the main results of the manuscript.

2.3.1 Assumptions

To establish the result the following sets of assumptions for the underlying stochastic sequence, the kernel function, density functions, as well as the form of dependence, among others is used.

Assumptions 2.3.1.

a) The sequence $\{X_i\}_{i=1}^{n+1}$ of random variables is strictly stationary.

b) The response function $\phi(\cdot)$ is Borel measurable on the real line with finite $\mathbb{E}|\phi(X_{i+1})|^s$ for some $s > 2$.

c) The derivatives $m'(x)$ and $m''(x)$ exist, are bounded, and uniformly continuous on $\mathbb{R}$.

The stationarity assumption is necessary because not many non-stationary processes have been identified to be strongly mixing. However, in some applications, such as the time-varying ARCH model, this assumption would indeed not be appropriate (see Fryzlewicz and Subba Rao (2011)). Considering a weaker assumption is possible but comes at an enormous notational expense in the proofs without additional insights. For uniform convergence of kernel estimators with heterogeneous dependent data see Kristensen (2009). The moment condition, Assumption 2.3.1.b), is specific to the model being analyzed. For example, when estimating the conditional CDF, the response function is bounded by one and therefore $\mathbb{E}|\phi(X_{i+1})|^s < \infty$ for all $s \geq 1$. The assumption is sufficiently weak but implies that responses, $\phi(X_{i+1})$, are not necessarily bounded. This provokes a truncation argument later in the manuscript. The last assumption is essential due to the construction of the estimator by approximating the unknown regression function locally by a polynomial of order $q + 1$.

For the general case assume $\{(X_i, Y_i)\}_{i=1}^{n}$ to be strictly stationary. All other assumption are similar to the once considered above by setting $X_{i+1} = Y_i$. This remains true for the following two assumptions.

Assumptions 2.3.2.

a) The kernel function $K(\cdot)$ is a symmetric and bounded density, i.e, for all $u \in \mathbb{R}$ $K(u) = K(-u)$ and $\sup_{u\in\mathbb{R}} K(u) \leq C_1 < \infty$.

b) $K(\cdot)$ has compact support $[-1, 1]$, i.e., $K(u) = 0$ for $|u| \geq 1$.

c) The kernel function $K(\cdot)$ is Lipschitz continuous, i.e., $|K(u) - K(v)| \leq C_2|u - v|$.

d) The function $u^j K^j(u)$ is Lipschitz continuous, i.e., $|u^j K^j(u) - v^j K^j(v)| \leq C_3|u - v|$ for all $j = 1, 2$.

Assumption 2.3.2.a) and 2.3.2.b) are common in the literature. Symmetry of the kernel function implies that the weighting scheme depends only on the absolute distance between the observation X_i and the evaluation point x. Fixing the support of the kernel function to $[-1, 1]$ is without loss of generality. However, finite support excludes kernel functions such as the Gaussian kernel. It is possible to mitigate this assumption by controlling the tail behavior of the kernel function. However, due to Fan et al. (1995), the Epanechnikov kernel, fulfilling the above assumption, is optimal in a minimax sense for the local linear estimator supporting this particular assumption. A consequence of the first two assumptions is that $\sup_{u\in\mathbb{R}} |u|^j K^j(u) = C_4$ for $j = 1, 2$ and $\int |u|K(u)du \leq C_5$ with finite strictly positive constants C_4 and C_5 (a proof can be found in Steikert (2014b)).

Assumptions 2.3.3.

a) The marginal density of X_i, $f_{X_i}(\cdot)$, is bounded by $C_6 > 0$.

b) The density $f_{X_i}(\cdot)$ is Lipschitz continuous, i.e., $|f_{X_i}(x_1) - f_{X_i}(x_2)| \leq C_7|x_1 - x_2|$, with $C_7 > 0$.

c) For compact subset $\mathbb{S} \subset \mathbb{R}$ the density $f_{X_i}(\cdot)$ is uniformly bounded from below on $\mathbb{S}$. That is, $\inf_{x\in\mathbb{S}} f_{X_i}(x) = C_8$, with $C_8 > 0$.

d) The joint density $f_{X_1,X_i}(x_1, x_i)$ of (X_1, X_i), with $i > 1$, is bounded by $C_9 > 0$.

e) The conditional density $f_{X_1,X_i|X_2,X_{i+1}}(x_1, x_i|x_2, x_{i+1})$ of (X_i, X_1) given (X_{i+1}, X_2), with $i > 1$, is bounded by $C_{10} > 0$.

Assumption 2.3.3.c) is needed for certain terms to be well-defined on the compact interval $\mathbb{S}$. The assumptions thereafter are common in the literature and are needed to bound various covariance terms emerging later in the text.

Because data is assumed to be weakly dependent, in particular a univariate time series framework is considered, the specific form of dependence needs to be specified. For this let $\mathcal{B}_s^t = \sigma\{X_i : s \leq i \leq t\}$ denote the σ-algebra generated by the time

series segment $\{X_s, X_{s+1}, ..., X_t\}$. The α- or strong-mixing coefficient, introduced by Rosenblatt (1956), is defined as

$$\alpha(k) = \sup_{A \in \mathcal{B}_{-\infty}^{0}, B \in \mathcal{B}_{k}^{\infty}} \left| \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \right|. \tag{2.9}$$

The mixing coefficient is therefore the total variation distance between two distributions. It measures the closeness of the joint distribution and product of the marginal distributions of the entire past and entire future k lags from today. If $\alpha(k) \to 0^+$ as $k \to \infty$ the joint distribution and the product of marginal distributions are arbitrary close implying asymptotic independence of the past and future of the sequence $\{X_i\}_{i=-\infty}^{\infty}$. Sequences fulfilling this condition are called strongly mixing sequences.

Assumption 2.3.4. The underlying stochastic sequence, $\{X_i\}_{i=1}^{n+1}$, is strongly mixing with mixing coefficient $\alpha(k)$, given in (2.9), satisfying $\sum_{k=1}^{\infty} k^a (\alpha(k))^{1-2/\delta} < \infty$ for some $\delta > 2$, with $a > 1 - 2/\delta$.

The δ in Assumption 2.3.4 refers to the order of the moment $\mathbb{E}\,|X_i|^\delta < \infty$ and is therefore specific to the underlying stochastic sequence under study. The strong mixing coefficient must guarantee summability of the stated sum. To adjust Assumption 2.3.4 to the general case define $\mathcal{B}_s^t = \sigma\{(X_i, Y_i) : s \leq i \leq t\}$ and the strong mixing coefficient similar as is (2.9).

To specify a rate of decay satisfying Assumption 2.3.4, assuming a polynomial strong mixing coefficient, the following lemma is of practical interest.

Lemma 2.3.5. *Suppose at least a polynomial strong mixing coefficient with rate of decay denoted by $\beta > 0$, i.e., $\alpha(k) \leq Ck^{-\beta}$ with nonnegative constant C, then the process $\{X_i\}_{i=1}^{n+1}$ satisfies Assumption 2.3.4 if*

$$\beta > \frac{2(\delta - 1)}{\delta - 2}. \tag{2.10}$$

To provide some intuition for the above lemma consider the following example. Let $\mathbb{E}\,|X_i|^\delta < \infty$ for all $\delta \geq 1$, then all moments of X_i exist and therefore Assumption 2.3.4 is fulfilled as long as $\beta > 2$. Hence, the fewer moments of X_i exist, the larger the rate of decay must be to fulfill Assumption 2.3.4.

2.3.2 Main results

Because a closed form solution of $\lambda_n(x)$ does not exist assessing the probabilities $p_i(x; \lambda_n)$ is not straightforward. I therefore prove in Theorem 2.3.7 that for any

compact subset $\mathbb{S}$ of $\mathbb{R}$ the probabilities $p_i(x;\lambda_n)$ are uniformly close to n^{-1} with probability one. Hence asymptotically the probabilities do not deviate from uniformity, i.e., n^{-1}, on $\mathbb{S}$ to fulfill condition (2.6). To establish the theorem define

$$S_{n,j}(x) = \frac{1}{nh_n}\sum_{i=1}^{n}(X_i - x)^j K_{h_n}^j(X_i - x), \tag{2.11}$$

for $j = 1, 2$. The partial sum given in (2.11), or variants thereof, are common objects when studying consistency problems of nonparametric estimators (see, e.g., Fan and Gijbels (1996, Section 3) and Masry (1996, p. 573)). The following lemma establishes strong uniform convergence of the partial sum $S_{n,j}(x)$ on $\mathbb{S}$. To prove the result the exponential type inequality by Liebscher (1996) given in Lemma B.1 in Appendix B is used. For the lemma let $\lfloor x \rfloor$ be the integer part of the real number x.

Lemma 2.3.6. *Suppose $h_n \to 0$, $nh_n/\ln(n) \to \infty$, as $n \to \infty$, and that Assumptions 2.3.1–2.3.4 hold. Given $S_{n,j}(x)$, defined in* (2.11)*, if*

i)

$$\Theta_n = \frac{n}{h_n^2}\alpha\left(\left\lfloor\sqrt{\frac{nh_n}{\ln(n)}}\right\rfloor\right),$$

is summable, i.e., $\sum_{n=1}^{\infty}\Theta_n < \infty$, then for any compact real interval $\mathbb{S}$,

$$\sup_{x\in\mathbb{S}}\left|S_{n,j}(x) - \nu_j f_{X_t}(x)\right| = \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \textit{almost surely}$$

for $j = 1, 2$ with $\nu_j = \int u^j K^j(u)du$.

ii) $\alpha(n) \le Cn^{-\beta}$ with β satisfying

$$\beta > \max\left\{4, \frac{2(\delta-1)}{\delta-2}\right\},$$

for some $\delta > 2$, and h_n such that $h_n \sim (\ln(n)/n)^\theta$ with

$$\theta \in \left(0, \frac{\beta-4}{\beta+4}\right),$$

then, for any compact real interval $\mathbb{S}$*,*

$$\sup_{x\in\mathbb{S}}\left|S_{n,j}(x)-\nu_j f_{X_t}(x)\right| = \begin{cases} \mathcal{O}\left(\left(\frac{\ln(n)}{n}\right)^{\theta}\right) & \text{almost surely,} \quad \text{if } \theta\in\left(0,\tfrac{1}{3}\right); \\ \mathcal{O}\left(\left(\frac{\ln(n)}{n}\right)^{(1-\theta)/2}\right) & \text{almost surely,} \quad \text{if } \theta\in\left[\tfrac{1}{3},1\right). \end{cases}$$

In the first part of Lemma 2.3.6 the condition regarding the α-mixing coefficient is weaker than in Masry (1996).[3] Assuming $h_n \sim (\ln(n)/n)^\theta$ in the second part does not contradict $\ln(n)/(nh_n) \to 0$. However, for low values of θ the bandwidth converges to zero slower than for high values. For an illustration suppose the order, δ, of the moment condition $\mathbb{E}\left|X_i\right|^\delta < \infty$ satisfies $\delta > 2$ implying $\beta > 4$. Then, if β is close to 4, θ is close to 0 which implies slow convergence. The larger θ the faster the bandwidth converges and therefore the larger the value of the rate of decay must be.

The next theorem establishes uniform convergence of $\lambda_n(x)$ on compact intervals $\mathbb{S}$. In addition, I establish that asymptotically the probabilities are selected equally.

Theorem 2.3.7. *Suppose that Assumptions 2.3.1–2.3.4 and condition i) of Lemma 2.3.6 hold. Then for any compact real interval* $\mathbb{S}$*,*

$$\sup_{x\in\mathbb{S}}\left|\lambda_n(x)\right| = \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \text{almost surely,}$$

and

$$\sup_{x\in\mathbb{S}}\left|p_i(x;\lambda_n)-\frac{1}{n}\right| = \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \text{almost surely.}$$

Note that a similar result as in the second part of Lemma 2.3.6 is easily established. However, to avoid redundancies I relinquish to do so until Lemma 2.3.14 but emphasize that the results are similar. Intuitively Theorem 2.3.7 states that, uniformly for $x \in \mathbb{S}$, the probabilities are selected equally for all i as $n \to \infty$. The estimator, however, remains constrained and therefore honors conditions (2.5) and (2.6) leading to the desirable bias property while being a proper estimator of the conditional CDF.

[3] In particular see Masry (1996, Proposition 1 and Theorem 1) for a similar result but note the difference in the definition of the partial sum $S_{n,j}(x)$.

Before proving uniform strong consistency for $\widehat{m}(x)$, given in (2.4), I rewrite the consistency problem as follows

$$\begin{aligned}\sup_{x\in\mathbb{S}}\left|\widehat{m}(x)-m(x)\right| &= \sup_{x\in\mathbb{S}}\left|\frac{\sum_{i=1}^{n}(\phi(X_{i+1})-m(x))p_i(x;\lambda_n)K_{h_n}(X_i-x)}{\sum_{i=1}^{n}p_i(x;\lambda_n)K_{h_n}(X_i-x)}\right| \\ &= \sup_{x\in\mathbb{S}}\left|\frac{\sum_{i=1}^{n}(\phi(X_{i+1})-m(X_i))p_i(x;\lambda_n)K_{h_n}(X_i-x)}{\sum_{i=1}^{n}p_i(x;\lambda_n)K_{h_n}(X_i-x)}\right. \\ &\qquad \left.+\frac{\sum_{i=1}^{n}(m(X_i)-m(x))p_i(x;\lambda_n)K_{h_n}(X_i-x)}{\sum_{i=1}^{n}p_i(x;\lambda_n)K_{h_n}(X_i-x)}\right| \\ &= \sup_{x\in\mathbb{S}}\left|\frac{J_1(x)}{J_3(x)}+\frac{J_2(x)}{J_3(x)}\right|, \end{aligned} \tag{2.12}$$

with

$$J_3(x)=\sum_{i=1}^{n}p_i(x;\lambda_n)K_{h_n}(X_i-x), \tag{2.13}$$

$$J_2(x)=\sum_{i=1}^{n}(m(X_i)-m(x))p_i(x;\lambda_n)K_{h_n}(X_i-x), \tag{2.14}$$

$$J_1(x)=\sum_{i=1}^{n}(\phi(X_{i+1})-m(X_i))p_i(x;\lambda_n)K_{h_n}(X_i-x). \tag{2.15}$$

Proving uniform strong consistency of $\widehat{m}(x)$ implies proving the uniform asymptotic behavior on $\mathbb{S}$ for each of the expressions in (2.13)–(2.15). I establish this separately in Propositions 2.3.8, 2.3.10, and 2.3.15. The corollaries following the first two propositions establish the uniform asymptotic behavior of expressions such as $\sup_{x\in\mathbb{S}}|J_3^{-1}(x)|$ and $\sup_{x\in\mathbb{S}}|J_2(x)|$ needed to prove the main result in Theorem 2.3.16.

The following proposition establishes that the uniform norm of the deviation $J_3(x)-f_{X_i}(x)$ on compact sets $\mathbb{S}\subset\mathbb{R}$ converges to zero.

Proposition 2.3.8. *Suppose that Assumptions 2.3.1–2.3.4 and condition i) of Lemma 2.3.6 hold. Then for any compact real interval* $\mathbb{S}$ *and* $J_3(x)$ *defined in* (2.13),

$$\sup_{x\in\mathbb{S}}\left|J_3(x)-f_{X_i}(x)\right|=\mathcal{O}(h_n)+\mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \textit{almost surely.}$$

As it may have been conjectured from Theorem 2.3.7, the constrained kernel density estimator $J_3(x)$ behaves similar as the ordinary kernel density estimator.

Note that the established rate of convergence is slightly slower than for the similar pointwise case (see Steikert (2014b, Proposition 1)). This is owed to the lack of continuity of the marginal density $f_{X_i}(\cdot)$ in the present framework. The following corollary is a straightforward implication of Proposition 2.3.8.

Corollary 2.3.9. *Given the assumptions of Proposition 2.3.8 and $J_3(x)$ defined in* (2.13)*, then for any compact real interval $\mathbb{S}$,*

$$\sup_{x\in\mathbb{S}}\left|\frac{1}{J_3(x)}\right| \leq \frac{1}{C_8}\left(1+\mathcal{O}(h_n)+\mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right)\right) \qquad \textit{almost surely,}$$

with C_8 defined in Assumption 2.3.3.c).

Next I establish the result for $J_2(x)$. This expression, given in (2.14), represents part of the bias of the estimator because it is the weighted difference of $m(X_i)$ and a zeroth-order approximation thereof.

Proposition 2.3.10. *Suppose that Assumptions 2.3.1–2.3.4 and condition i) of Lemma 2.3.6 hold. Then for any compact real interval $\mathbb{S}$ and $J_2(x)$ defined in* (2.14)*,*

$$\sup_{x\in\mathbb{S}}\left|J_2(x)-\frac{h_n^2}{2}\mu_1 m''(x)f_{X_i}(x)\right| = o(h_n^2) \qquad \textit{almost surely,}$$

with $\mu_1 = \int u^2K(u)du$.

Corollary 2.3.11. *Given the assumptions of Proposition 2.3.10 and $J_2(x)$ defined in* (2.14)*, then for any compact real interval $\mathbb{S}$,*

$$\sup_{x\in\mathbb{S}}\left|J_2(x)\right| = \mathcal{O}(h_n^2) \qquad \textit{almost surely.}$$

The expression $J_1(x)$, given in (2.15), represents the weighted error of the data generating process. To establish its strong uniform asymptotic behavior, the well-known truncation argument introduced in Mack and Silverman (1982, p. 408) is employed. This is because $\phi(X_{i+1})$ is not necessarily bounded which is owed to Assumption 2.3.1.b). An alternative is to assume a bounded response function. This assumption, however, is rather strong.

Lemma 2.3.12. *Given Assumption 2.3.1 and*

$$\tau_n = \big(n(\ln(\ln(n)))^2 \ln(n)\big)^{1/s}, \tag{2.16}$$

for some $s > 2$, then

$$\big|\phi(X_{i+1})\big| \leq \tau_n \qquad \textit{almost surely,}$$

for $i \leq n$ and n sufficiently large.

Note that τ_n depends on the order s of the moment $\mathbb{E}\,|\phi(X_{i+1})|^s < \infty$. That is, τ_n increases faster in n for lower rather than for higher values of s. In the case of $\phi(X_{i+1}) = \mathbb{1}_{(-\infty,y]}(X_{i+1})$ for some $y \in \mathbb{R}$ infinite moments exist and therefore $\tau_n = 1$ for all n which is an obvious choice to bound an indicator function.

Similar to $S_{n,j}(x)$, defined in (2.11), the partial sum $T_n(x)$ is defined as

$$T_n(x) = \frac{1}{n}\sum_{i=1}^{n}(\phi(X_{i+1}) - m(X_i))K_{h_n}(X_i - x), \tag{2.17}$$

where it is easy to see that $\mathbb{E}\,T_n(x) = 0$. Using $T_n(x)$, however, is not appropriate in this setting because the response function is not necessarily bounded. Consider instead the truncation of $T_n(x)$, given the bound τ_n, which reads

$$T_n^{(t)}(x) = \frac{1}{n}\sum_{i=1}^{n}\big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}} - m^{(t)}(X_i)\big)K_{h_n}(X_i - x), \tag{2.18}$$

with

$$m^{(t)}(X_i) = \mathbb{E}\big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}|X_i\big), \tag{2.19}$$

denoting the truncated version of $m(X_i)$. The next corollary provides reason to work with the truncation (2.18) instead of (2.17).

Corollary 2.3.13. *Given the assumptions of Lemma 2.3.12 and*

$$\begin{aligned} T_n(x) - T_n^{(t)}(x) = \frac{1}{n}\sum_{i=1}^{n}\big(&\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|>\tau_n\}} \\ &- \mathbb{E}\big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|>\tau_n\}}|X_i\big)\big)K_{h_n}(X_i - x), \end{aligned}$$

then

$$T_n(x) - T_n^{(t)}(x) = o(1) \qquad \textit{almost surely.}$$

Thus, replacing $T_n(x)$ with $T_n^{(t)}(x)$ results in an error tending to zero as $n \to \infty$ with probability one. For the remainder of this section I henceforth work with the truncation and establish its uniform strong consistency in the next lemma.

Lemma 2.3.14. *Suppose $h_n \to 0$, $nh_n/\ln(n) \to \infty$, as $n \to \infty$, $s \geq \delta > 2$, and that Assumptions 2.3.1–2.3.4 hold. Given $T_n^{(t)}(x)$ and τ_n, defined in* (2.18) *and* (2.16)*, respectively, if*

i)

$$\Xi_n = n\left(\frac{\tau_n}{h_n}\right)^2 \alpha\left(\frac{1}{\tau_n}\left\lfloor\sqrt{\frac{nh_n}{\ln(n)}}\right\rfloor\right),$$

is summable, i.e., $\sum_{n=1}^{\infty} \Xi_n < \infty$, then for any compact real interval $\mathbb{S}$,

$$\sup_{x\in\mathbb{S}} \left| T_n^{(t)}(x)\right| = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \textit{almost surely.}$$

ii) $\alpha(n) \leq Cn^{-\beta}$ with the rate of decay satisfying

$$\beta > \frac{4(s+1)}{s-2},$$

and h_n such that $h_n \sim (\ln(n)/n)^{\theta}$ with

$$\theta \in \left(0, \frac{\beta(1-2/s) - 4/s - 4}{\beta + 4}\right),$$

then for any compact real interval $\mathbb{S}$,

$$\sup_{x\in\mathbb{S}} \left| T_n^{(t)}(x)\right| = \mathcal{O}\left(\left(\frac{\ln(n)}{nh_n}\right)^{(1-\theta)/2}\right) \quad \textit{almost surely.}$$

Compared to Lemma 2.3.6 the assumption regarding the α-mixing coefficient and the rate of decay, β, are stronger. This is due to the truncation argument made above and is seen in condition i) by the argument of the strong mixing coefficient

and the multiplication of τ_n^2. For an illustration of condition ii) consider again the response function $\phi(X_{i+1}) = \mathbb{1}_{(-\infty,y]}(X_{i+1})$ for some $y \in \mathbb{R}$, then $\mathbb{E}\,|\phi(X_{i+1})|^s < \infty$ for all $s \geq 1$. If, e.g., $\theta = 1/5$, then $\beta > 6$. It can therefore easily be seen that there is a tradeoff between s and β, implying the need for a higher rate of decay for fewer existing moments.

The next proposition establishes uniform strong consistency of $J_1(x)$ on $\mathbb{S}$.

Proposition 2.3.15. *Suppose that Assumptions 2.3.1–2.3.4 and condition i) of Lemma 2.3.14 hold. Then for any compact real interval $\mathbb{S}$ and $J_1(x)$ defined in* (2.15)*,*

$$\sup_{x\in\mathbb{S}} \left|J_1(x)\right| = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \textit{almost surely.}$$

Corollaries 2.3.9 and 2.3.11 as well as Proposition 2.3.15 enable to prove uniform strong consistency on compact subsets $\mathbb{S}$ for the weighted Nadaraya-Watson estimator.

Theorem 2.3.16. *Suppose that Assumptions 2.3.1–2.3.4 hold. Given the weighted Nadaraya-Watson estimator $\widehat{m}(x)$, defined in* (2.4)*, and*

i) that condition i) of Lemma 2.3.14 holds, then for any compact real interval $\mathbb{S}$,

$$\sup_{x\in\mathbb{S}} \left|\widehat{m}(x) - m(x)\right| = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) + \mathcal{O}(h_n^2) \qquad \textit{almost surely.}$$

ii) that condition ii) of Lemma 2.3.14 holds, then for any compact real interval $\mathbb{S}$,

$$\sup_{x\in\mathbb{S}} \left|\widehat{m}(x) - m(x)\right| = \begin{cases} \mathcal{O}\left(\left(\frac{\ln(n)}{n}\right)^{2\theta}\right) & \textit{almost surely, if } \theta \in \left(0, \frac{1}{5}\right); \\ \mathcal{O}\left(\left(\frac{\ln(n)}{n}\right)^{(1-\theta)/2}\right) & \textit{almost surely, if } \theta \in \left[\frac{1}{5}, 1\right). \end{cases}$$

Note that the assumptions of Lemma 2.3.14 are used because these are stronger than those of Lemma 2.3.6. The rate of convergence in the first part of the above theorem is the same as for the local linear estimator (see Masry 1996, Theorem 6). Note also that the rate established in Theorem 2.3.16 is the same as for the case of

pointwise strong consistency (see Steikert 2014b, Theorem 2). Regarding the second part, note that if $h_n \sim (\ln(n)/n)^{1/5}$, i.e., $\theta = 1/5$, then the convergence rate is of order $(\ln(n)/n)^{2/5}$. This rate is optimal in the sense of Stone (1982, Theorem 1) for independent and identically distributed data.

2.4 Conclusion

In this manuscript uniform strong consistency for the weighted Nadaraya-Watson estimator for strongly mixing processes is established. In addition, assuming a polynomial strong mixing coefficient, necessary conditions on the rate of decay depending on the speed of convergence of the bandwidth are established for a more practical perspective regarding the rate of convergence. The established rate is optimal and equivalent to the strong rates of convergence of the local linear estimator. The uniform strong consistency result is fundamental for further research of estimators in which the weighted Nadaraya-Watson estimator is embedded, in particular for proving consistency results. Examples of such estimators are two-stage, semiparametric, or bootstrap estimators.

Acknowledgements

I thank Simon Broda, Silvia Grätz, Michael Wolf, and Jan Wrampelmeyer for valuable comments and suggestions.

Appendices

A Proofs

The constant $\overline{C}$ in the proofs below is suitable for each expression. That is, it represents (possibly) different values at different places. For proving the general case one may substitute $X_{i+1} = Y_i$ in the proofs below and use the set of generalized assumptions.

A.1 Proof of Lemma 2.3.5

Suppose $\alpha(k) \leq Ck^{-\beta}$, then, given Assumption 2.3.4,

$$\sum_{k=1}^{\infty} k^a (\alpha(k))^{1-2/\delta} \leq C \sum_{k=1}^{\infty} k^{a-\beta(1-2/\delta)}.$$

For the infinite sum to convergence it must be that $a - \beta(1 - 2/\delta) < -1$. Solving this expression for $\beta > 0$, given $a > 1 - 2/\delta$, proves the statement.

□

A.2 Proof of Lemma 2.3.6

For the first statement I separate the problem by adding and subtracting $\mathbb{E}\, S_{n,j}(x)$, this leads to

$$\begin{aligned} \sup_{x\in\mathbb{S}} \left|S_{n,j}(x) - \nu_j f_{X_i}(x)\right| &\leq \sup_{x\in\mathbb{S}} \left\{ \left| \mathbb{E}\, S_{n,j}(x) - \nu_j f_{X_i}(x) \right| \right\} + \sup_{x\in\mathbb{S}} \left\{ \left| S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x) \right| \right\} \\ &=: A_n + B_n. \end{aligned} \tag{A.1}$$

Even for $x \in \mathbb{R}$ the non-stochastic term, i.e., the first summand of (A.1), can be bounded as follows

$$\begin{aligned} A_n &= \sup_{x\in\mathbb{R}} \left| \frac{1}{h_n} \mathbb{E}(X_1 - x)^j K_{h_n}^j (X_1 - x) - \nu_j f_{X_i}(x) \right| \\ &= \sup_{x\in\mathbb{R}} \left| \frac{1}{h_n} \int \left(\frac{z-x}{h_n} \right)^j K^j \left(\frac{z-x}{h_n} \right) f_{X_i}(z) dz - \nu_j f_{X_i}(x) \right| \end{aligned}$$

$$
\begin{aligned}
&= \sup_{x\in\mathbb{R}} \left| \int u^j K^j(u) f_{X_i}(x+uh_n)dz - \nu_j f_{X_i}(x) \right| \\
&= \sup_{x\in\mathbb{R}} \left| \int u^j K^j(u)(f_{X_i}(x+uh_n) - f_{X_i}(x))du \right| \\
&\leq \sup_{x\in\mathbb{R}} \int |u|^j K^j(u) \big| f_{X_i}(x+uh_n) - f_{X_i}(x) \big| du \\
&\leq h_n C_7 \int |u|^{j+1} K^j(u) du \\
&= O(h_n). \qquad \text{(A.2)}
\end{aligned}
$$

The first line follows from the definition of $S_{n,j}(x)$, given in (2.11), and stationarity. The third and fourth uses a change of variable, i.e., $u = (z-x)/h_n$ and the definition of ν_j, respectively. The second to last line is due to the assumption that $f_{X_i}(\cdot)$ is Lipschitz continuous. The integral in the last line is bounded because $K(\cdot)$ is bounded and has bounded support.

For B_n, the second term of (A.1), establishing

$$
\sup_{x\in\mathbb{S}} \big| S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x) \big| = O\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \qquad \text{almost surely,}
$$

proves the first part of the lemma. Because $\mathbb{S}$ is a compact real interval it can be covered with a finite number, L_n, of subintervals I_k with centers x_k and length l_n, where $k = 1, ..., L_n$. Then,

$$
\begin{aligned}
B_n &= \sup_{x\in\mathbb{S}} \big| S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x) \big| \\
&= \max_{1\leq k\leq L_n} \sup_{x\in\mathbb{S}\cap I_k} \big| S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x) \big| \\
&= \max_{1\leq k\leq L_n} \sup_{x\in\mathbb{S}\cap I_k} \big| S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x) + S_{n,j}(x_k) - \mathbb{E}\, S_{n,j}(x_k) - (S_{n,j}(x_k) - \mathbb{E}\, S_{n,j}(x_k)) \big| \\
&\leq \max_{1\leq k\leq L_n} \left\{ \sup_{x\in\mathbb{S}\cap I_k} \big| S_{n,j}(x) - S_{n,j}(x_k) \big| \right\} + \max_{1\leq k\leq L_n} \big| S_{n,j}(x_k) - \mathbb{E}\, S_{n,j}(x_k) \big| \\
&\qquad + \max_{1\leq k\leq L_n} \sup_{x\in\mathbb{S}\cap I_k} \big| \mathbb{E}\, S_{n,j}(x_k) - \mathbb{E}\, S_{n,j}(x) \big| \\
&=: B_{n,1} + B_{n,2} + B_{n,3}, \qquad \text{(A.3)}
\end{aligned}
$$

For $B_{n,1}$ it is easy to see, using the definition of $S_{n,j}(x)$ in (2.11), that

$$\begin{aligned}
B_{n,1} &= \max_{1\le k\le L_n} \sup_{x\in\mathbb{S}\cap I_k} \left| \frac{1}{nh_n} \sum_{i=1}^{n} \Big((X_i - x)^j K_{h_n}^j (X_i - x) - (X_i - x_k)^j K_{h_n}^j (X_i - x_k) \Big) \right| \\
&\le \max_{1\le k\le L_n} \sup_{x\in\mathbb{S}\cap I_k} \left\{ \frac{1}{nh_n} \sum_{i=1}^{n} \left| (X_i - x)^j K_{h_n}^j (X_i - x) - (X_i - x_k)^j K_{h_n}^j (X_i - x_k) \right| \right\} \\
&\le \max_{1\le k\le L_n} \sup_{x\in\mathbb{S}\cap I_k} \left\{ \frac{C_3}{h_n} \left| \frac{x_k - x}{h_n} \right| \right\} \\
&\le \frac{C_3 l_n}{h_n^2} \\
&= \frac{\overline{C}}{h_n^2 L_n} \\
&\le \overline{C} \sqrt{\frac{\ln(n)}{nh_n}}.
\end{aligned}$$

The third line follows from Assumption 2.3.2.d), the fourth from the fact that the maximal distance between x_k and x is equivalent to the length of the subinterval l_n, and the fifth from the definition of the length l_n, i.e., $l_n = \overline{C}/L_n$ with suitable constant $\overline{C}$ (here diameter of $\mathbb{S}$). For the last line set the number of subintervals I_k to $L_n = \lceil \sqrt{n/(h_n^3 \ln(n))} \rceil$, with $\lceil x \rceil$ being the smallest integer m such that $m \ge x$, and use the fact that $\sqrt{n/(h_n^3 \ln(n))} \le L_n$. It is easily seen that $\mathbb{P}(B_{n,1} \le \overline{C}\sqrt{\ln(n)/(nh_n)}) = 1$ and therefore

$$B_{n,1} = O\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \quad \text{almost surely.} \tag{A.4}$$

For $B_{n,3}$, the third term of (A.3), similar arguments as for $B_{n,1}$ lead to

$$B_{n,3} = O\left(\sqrt{\frac{\ln(n)}{nh_n}} \right). \tag{A.5}$$

Regarding $B_{n,2}$ the Borel-Cantelli lemma is used to prove almost sure convergence, showing that for each $\varepsilon > 0$, $\sum_{n=1}^{\infty} \mathbb{P}(B_{n,2} > \varepsilon) < \infty$. Rewriting the problem leads to

$$\begin{aligned}
\mathbb{P}(B_{n,2} > \varepsilon) &= \mathbb{P}\left(\max_{1\le k\le L_n} \left| S_{n,j}(x_k) - \mathbb{E}\, S_{n,j}(x_k) \right| > \varepsilon \right) \\
&\le \sum_{k=1}^{L_n} \mathbb{P}\left(\left| S_{n,j}(x_k) - \mathbb{E}\, S_{n,j}(x_k) \right| > \varepsilon \right) \\
&\le L_n \sup_{x\in\mathbb{R}} \left\{ \mathbb{P}\left(\left| S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x) \right| > \varepsilon \right) \right\}.
\end{aligned} \tag{A.6}$$

To bound the probability term of (A.6) the exponential type inequality of Lemma B.1 given in Appendix B is used. For this define

$$Z_{j,i}(x) = (X_i - x)^j K_{h_n}^j(X_i - x) - \mathbb{E}(X_1 - x)^j K_{h_n}^j(X_1 - x), \tag{A.7}$$

for $j = 1, 2$, implying

$$\mathbb{P}\big(\big|S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x)\big| > \varepsilon\big) = \mathbb{P}\left(\Big|\sum_{i=1}^{n} Z_{j,i}(x)\Big| > \varepsilon n h_n\right).$$

To apply Lemma B.1 to the above expression the bound b, the variance $\sigma_{s_n}^2$ (defined in (B.82)), and the integer $s_n \leq n$ such that $s_n < \varepsilon n h_n b/4$ need to be established. For b it is easy to see that

$$\begin{aligned}
\big|Z_{j,i}(x)\big| &= \Big|(X_i - x)^j K_{h_n}^j(X_i - x) - \mathbb{E}(X_1 - x)^j K_{h_n}^j(X_1 - x)\Big| \\
&\leq \big|X_i - x\big|^j K_{h_n}^j(X_i - x) + \mathbb{E}\,\big|X_1 - x\big|^j K_{h_n}^j(X_1 - x)\Big| \\
&\leq 2C_4 =: b,
\end{aligned}$$

since $\sup_{u\in\mathbb{R}} |u|^j K^j(u)du = C_4$ and $K(\cdot)$ being bounded having bounded support.

Before determining $\sigma_{s_n}^2$, with s_n not yet defined but fulfilling $s_n \leq n$, note that

$$\begin{aligned}
\mathbb{E}(X_1 - x)^j K_{h_n}^j(X_1 - x) &= \int (z - x)^j K_{h_n}^j(z - x) f_{X_i}(z) dz \\
&= h_n \int u^j K^j(u) f_{X_i}(x + u h_n) du \\
&\leq C_6 h_n \int |u|^j K^j(u) du \\
&= \mathcal{O}(h_n),
\end{aligned} \tag{A.8}$$

using a change of variable, i.e., $u = (z-x)/h_n$ and the boundedness of the marginal density given in Assumption 2.3.3.a). The last line follows from $K(\cdot)$ being bounded and having bounded support. Rewriting $\sigma_{s_n}^2$, defined in (B.82), leads to

$$\begin{aligned}
\sigma_{s_n}^2 &= \mathbb{E}\left(\sum_{i=1}^{s_n} Z_{j,i}(x)\right)^2 \\
&= \sum_{i=1}^{s_n} \mathbb{E}\, Z_{j,i}^2(x) + \sum_{i=1}^{s_n} \sum_{\substack{k=1 \\ k\neq i}}^{s_n} \mathbb{E}\, Z_{j,i}(x) Z_{j,k}(x)
\end{aligned}$$

$$
\begin{aligned}
&= s_n \,\mathbb{E}\, Z_{j,1}^2(x) + 2s_n \sum_{i=2}^{s_n} \left(1 - \frac{i-1}{s_n}\right) \mathbb{E}\, Z_{j,1}(x) Z_{j,i}(x) \\
&\leq s_n \,\mathbb{E}\, Z_{j,1}^2(x) + 2s_n \sum_{i=2}^{s_n} \left| \mathbb{E}\, Z_{j,1}(x) Z_{j,i}(x) \right| \\
&=: \Sigma_1 + \Sigma_2, \qquad \text{(A.9)}
\end{aligned}
$$

where the second line follows from stationarity.

To bound Σ_1 note that

$$
\begin{aligned}
\Sigma_1 &= s_n \,\mathbb{E}\, Z_{j,1}^2(x) \\
&= s_n \,\mathbb{E}\Big((X_1 - x)^j K_{h_n}^j (X_1 - x) - \mathbb{E}(X_1 - x)^j K_{h_n}^j (X_1 - x) \Big)^2 \\
&= s_n \left(\mathbb{E}\Big((X_1 - x)^j K_{h_n}^j (X_1 - x) \Big)^2 - \Big(\mathbb{E}(X_1 - x)^j K_{h_n}^j (X_1 - x) \Big)^2 \right) \\
&\leq s_n \left(\mathbb{E}\Big((X_1 - x)^j K_{h_n}^j (X_1 - x) \Big)^2 + \Big(\mathbb{E}(X_1 - x)^j K_{h_n}^j (X_1 - x) \Big)^2 \right) \\
&\leq s_n \Big(\mathbb{E}(X_1 - x)^{2j} K_{h_n}^{2j} (X_1 - x) + \mathcal{O}(h_n^2) \Big) \\
&= s_n \left(\int (z - x)^{2j} K_{h_n}^{2j} (z - x) f_{X_i}(z) dz + \mathcal{O}(h_n^2) \right) \\
&= s_n \left(h_n \int u^{2j} K^{2j}(u) f_{X_i}(x + u h_n) du + \mathcal{O}(h_n^2) \right) \\
&\leq s_n h_n (C_6 \nu_{2j} + \mathcal{O}(h_n)), \qquad \text{(A.10)}
\end{aligned}
$$

with $\nu_{2j} = \int u^{2j} K^{2j}(u) du$. The fourth line follows from (A.8) and the second to last from a change of variable, i.e., $u = (z - x)/h_n$. The last line is due to Assumption 2.3.3.a).

To bound Σ_2, defined in (A.9), note that

$$
\begin{aligned}
\Sigma_2 &= 2s_n \sum_{i=2}^{s_n} \left| \mathbb{E}\, Z_{j,1}(x) Z_{j,i}(x) \right| \\
&\leq 2s_n \sum_{i=2}^{\lfloor h_n^{-1} \rfloor} \left| \mathbb{E}\, Z_{j,1}(x) Z_{j,i}(x) \right| + 2s_n \sum_{i=\lfloor h_n^{-1} \rfloor + 1}^{\infty} \left| \mathbb{E}\, Z_{j,1}(x) Z_{j,i}(x) \right| \\
&=: \Sigma_{21} + \Sigma_{22}, \qquad \text{(A.11)}
\end{aligned}
$$

where $\lfloor x \rfloor$ denotes the integer part of $x \in \mathbb{R}$. Before bounding Σ_{21} and Σ_{22}, note that simple calculations reveal that

$$
\begin{aligned}
\mathbb{E}\, Z_{j,1}(x) Z_{j,i}(x) &= \operatorname{cov}(Z_{j,1}(x), Z_{j,i}(x)) \\
&= \operatorname{cov}\Big((X_1 - x)^j K_{h_n}^j (X_1 - x), (X_i - x)^j K_{h_n}^j (X_i - x) \Big).
\end{aligned}
$$

Bounding Σ_{21}, given in (A.11), may be done as follows

$$\begin{aligned}
\Sigma_{21} &= 2s_n \sum_{i=2}^{\lfloor h_n^{-1} \rfloor} \left| \mathbb{E}\, Z_{j,1}(x) Z_{j,i}(x) \right| \\
&= 2s_n \sum_{i=2}^{\lfloor h_n^{-1} \rfloor} \Big| \mathbb{E}(X_1 - x)^j K_{h_n}^j (X_1 - x)(X_i - x)^j K_{h_n}^j (X_i - x) \\
&\qquad\qquad - \mathbb{E}(X_1 - x)^j K_{h_n}^j (X_1 - x) \times \mathbb{E}(X_i - x)^j K_{h_n}^j (X_i - x) \Big| \\
&= 2s_n \sum_{i=2}^{\lfloor h_n^{-1} \rfloor} \left| \int (z - x)^j K_{h_n}^j (z - x)(w - x)^j K_{h_n}^j (w - x) \right. \\
&\qquad\qquad \left. \times \left(f_{X_1, X_i}(z, w) - f_{X_i}(z) f_{X_i}(w) \right) dz dw \right| \\
&\leq \text{constant}_1 \times s_n h_n^2 \lfloor h_n^{-1} \rfloor \int |u|^j K^j(u) |v|^j K^j(v) du dv \\
&\leq \text{constant}_1 \times s_n h_n \left(\int |u|^j K^j(u) du \right)^2, \qquad \text{(A.12)}
\end{aligned}$$

with constant_1 denoting a suitable constant. The second and third equation follows from the definition of the covariance. For the second to last line use two changes of variables, i.e., $u = (z - x)/h_n$ and $v = (w - x)/h_n$, and the fact that the marginal and joint densities are both bounded due to Assumption 2.3.3.a) and 2.3.3.d). The last line is owed to the fact that $\lfloor x \rfloor = x - z$ for any $x \in \mathbb{R}$ and some appropriate $z \in (0, 1)$.

To determine Σ_{22}, defined in (A.11), apply Davydov's lemma repeated in Lemma B.2 in Appendix B to each summand $|\, \mathbb{E}\, Z_{j,1}(x) Z_{j,i}(x)|$. To apply the lemma note that for $\delta > 2$,

$$\begin{aligned}
\mathbb{E} \left| (X_1 - x)^j K_{h_n}^j (X_1 - x) \right|^\delta &\leq \int |z - x|^{j\delta} K_{h_n}^{j\delta} (z - x) f_{X_i}(z) dz \\
&\leq h_n C_4^{j\delta - 1} C_5 C_6 \\
&= \overline{C} h_n \\
&< \infty,
\end{aligned}$$

and therefore

$$\left\| (X_1 - x)^j K_{h_n}^j (X_1 - x) \right\|_\delta^2 \leq \overline{C} h_n^{2/\delta},$$

with suitable constant $\overline{C}$. For $a > 1 - 2/\delta$ it follows, by virtue of Davydov's lemma, that

$$\begin{aligned}
\Sigma_{22} &= 2s_n \sum_{i=\lfloor h_n^{-1} \rfloor + 1}^{\infty} \left| \mathbb{E}\, Z_{j,1}(x) Z_{j,i}(x) \right| \\
&\leq \text{constant}_2 \times s_n h_n^{2/\delta} \sum_{i=\lfloor h_n^{-1} \rfloor + 1}^{\infty} (\alpha(i-1))^{1-2/\delta} \\
&= \text{constant}_2 \times s_n h_n^{2/\delta} \sum_{k=\lfloor h_n^{-1} \rfloor}^{\infty} (\alpha(k))^{1-2/\delta} \\
&\leq \text{constant}_2 \times s_n h_n^{2/\delta} \sum_{k=\lfloor h_n^{-1} \rfloor}^{\infty} \left(\frac{k}{\lfloor h_n^{-1} \rfloor} \right)^a (\alpha(k))^{1-2/\delta} \\
&\leq \text{constant}_2 \times s_n h_n \sum_{k=\lfloor h_n^{-1} \rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta}, \qquad \text{(A.13)}
\end{aligned}$$

with constant_2 denoting a suitable constant. The third line follows from a change of variable, i.e., $k = i - 1$. The fourth line is due to the fact that $(k/\lfloor h_n^{-1} \rfloor)^a \geq 1$ for $k = \lfloor h_n^{-1} \rfloor, \lfloor h_n^{-1} \rfloor + 1, \ldots$ and any $a > 0$. For the last line, in particular for the sum's multiplier, note that $\lfloor h_n^{-1} \rfloor^a = (h_n^{-1} - z)^a$, for some $z \in [0, 1)$. Then,

$$\begin{aligned}
\frac{h_n^{2/\delta}}{\lfloor h_n^{-1} \rfloor^a} &= \frac{h_n^{2/\delta}}{\left(h_n^{-1} - z\right)^a} \\
&= \frac{h_n^{a+2/\delta}}{(1 - z h_n)^a} \\
&\leq h_n^{a+2/\delta} \\
&\leq h_n,
\end{aligned}$$

because $a + 2/\delta > 1$, since $a > 1 - 2/\delta$, and $h_n \to 0$, as $n \to \infty$ and $1 - z h_n$ being asymptotically equivalent to 1. According to Assumption 2.3.4 it follows that for (A.13), $\sum_{k=\lfloor h_n^{-1} \rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta} = \text{o}(1)$ because $\lfloor h_n^{-1} \rfloor \to \infty$ as $n \to \infty$. Combining (A.9), (A.10), (A.11), (A.12), and (A.13) results in

$$\begin{aligned}
\sigma_{s_n}^2 \leq s_n h_n \Bigg(& C_6 \nu_{2j} + O(h_n) + \text{constant}_1 \times \left(\int |u|^j K^j(u) du \right)^2 \\
& + \text{constant}_2 \times \sum_{k=\lfloor h_n^{-1} \rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta} \Bigg). \qquad \text{(A.14)}
\end{aligned}$$

Given the bound b and the variance $\sigma_{s_n}^2$ it remains to prove $s_n < \eta b/4$ by defining $s_n = \lfloor \sqrt{n h_n / \ln(n)} \rfloor$, $\varepsilon = \varepsilon_n = C_\varepsilon = \sqrt{\ln(n)/(n h_n)}$, with $C_\varepsilon > 0$, and $\eta = \varepsilon_n n h_n$.

Thus,

$$\begin{aligned} s_n &< \frac{\eta b}{4} \\ \left\lfloor \sqrt{\frac{nh_n}{\ln(n)}} \right\rfloor &< \frac{C_4 C_\varepsilon nh_n}{2} \sqrt{\frac{\ln(n)}{nh_n}} \\ \iff \left(\sqrt{\frac{nh_n}{\ln(n)}} - z \right) \sqrt{\frac{nh_n}{\ln(n)}} &< \frac{C_4 C_\varepsilon nh_n}{2} \\ \iff 1 - z\sqrt{\frac{\ln(n)}{nh_n}} &< \frac{C_4 C_\varepsilon \ln(n)}{2}, \end{aligned}$$

for sufficiently large n. Applying Lemma B.1 leads to

$$\begin{aligned} \mathbb{P}\big(\big|S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x)\big| > \varepsilon_n\big) &= \mathbb{P}\left(\Big| \sum_{i=1}^{n} Z_{j,i}(x) \Big| > \varepsilon_n nh_n \right) \\ &\leq 4 \exp\left\{ -\frac{\varepsilon_n^2 (nh_n)^2}{64 \frac{n}{s_n} \sigma_{s_n}^2 + \frac{8}{3} \varepsilon_n nh_n s_n b} \right\} + 4 \frac{n}{s_n} \alpha(s_n). \end{aligned} \tag{A.15}$$

For the first term of (A.15) it follows that

$$\begin{aligned} 4 \exp\left\{ -\frac{\varepsilon_n^2 (nh_n)^2}{64 \frac{n}{s_n} \sigma_{s_n}^2 + \frac{8}{3} \varepsilon_n nh_n s_n b} \right\} &\leq 4 \exp\left\{ -\frac{C_\varepsilon^2 \ln(n) nh_n}{\overline{C} nh_n + \frac{16}{3} C_4 C_\varepsilon nh_n} \right\} \\ &= 4 \exp\left\{ -\frac{C_\varepsilon^2 \ln(n)}{\overline{C} + \frac{16}{3} C_4 C_\varepsilon} \right\} \\ &= 4 \exp\{-C_\varepsilon \ln(n)\}, \end{aligned} \tag{A.16}$$

with C_ε in the last line absorbing the constants from the previous line. Note that $\varepsilon_n \to 0$, as $n \to \infty$, for any nonnegative C_ε, hence it can be selected to fit the analysis best. To determine the second term of (A.15) note that for large n,

$$\frac{4n}{s_n} \alpha\left(\left\lfloor \sqrt{\frac{nh_n}{\ln(n)}} \right\rfloor \right) = 4 \sqrt{\frac{n \ln(n)}{h_n}} \alpha\left(\left\lfloor \sqrt{\frac{nh_n}{\ln(n)}} \right\rfloor \right), \tag{A.17}$$

because s_n and $\sqrt{\ln(n)/(nh_n)}$ are asymptotically equivalent. Combining (A.15), (A.16), and (A.17) with (A.6) leads to

$$L_n \sup_{x \in \mathbb{R}} \mathbb{P}\left(\big|S_{n,j}(x) - \mathbb{E}\, S_{n,j}(x)\big| > C_\varepsilon \sqrt{\frac{\ln(n)}{nh_n}} \right) \leq 4 L_n n^{-C_\varepsilon} + 4 L_n \sqrt{\frac{n \ln(n)}{h_n}} \alpha(s_n).$$

By the definition of L_n the first term is summable for C_ε sufficiently large. For the second

term note that

$$\begin{aligned} 4L_n\sqrt{\frac{n\ln(n)}{h_n}}\alpha(s_n) &= 4\left\lceil\sqrt{\frac{n}{h_n^3\ln(n)}}\right\rceil\sqrt{\frac{n\ln(n)}{h_n}}\alpha(s_n) \\ &= 4\left(\sqrt{\frac{n}{h_n^3\ln(n)}}+z\right)\sqrt{\frac{n\ln(n)}{h_n}}\alpha(s_n) \\ &= \frac{4n}{h_n^2}\alpha(s_n)+4z\sqrt{\frac{n\ln(n)}{h_n}}\alpha(s_n). \end{aligned} \tag{A.18}$$

The second line follows from $\lceil x\rceil = x+z$, for $x\in\mathbb{R}$ and some $z\in[0,1)$. Because the second summand is of smaller order than the first and the first is summable by condition i) of the lemma. Hence, $\sum_{n=1}^{\infty}\mathbb{P}(B_{n,2} > C_\varepsilon\sqrt{\ln(n)/(nh_n)}) < \infty$. By virtue of the Borel-Cantelli lemma it follows that

$$B_{n,2} = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \tag{A.19}$$

Combining (A.3), (A.4), (A.5), and (A.19) leads to

$$B_n = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \tag{A.20}$$

Finally, using (A.1), (A.2), and (A.20) results in

$$\sup_{x\in\mathbb{S}}\left|S_{n,j}(x)-\nu_j f_{X_i}(x)\right| = \mathcal{O}(h_n)+\mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely,} \tag{A.21}$$

for $j=1,2$. This completes the proof of the first statement.

For part ii) of the lemma use the first term of (A.18) with $\alpha(k)\le Ck^{-\beta}$ and $h_n\sim(\ln(n)/n)^\theta$, then

$$\begin{aligned} \frac{n}{h_n^2}\alpha(s_n) &= \frac{n}{h_n^2}\left(\frac{nh_n}{\ln(n)}\right)^{-\beta/2} \\ &= \frac{n^{1+2\theta}}{(\ln(n))^{2\theta}}\left(\frac{n}{\ln(n)}\right)^{-\beta(1-\theta)/2} \\ &= n^{1+2\theta-\beta(1-\theta)/2}(\ln(n))^{-2\theta+\beta(1-\theta)/2} \\ &= \ln(n)\left(\frac{n}{\ln(n)}\right)^{1+2\theta-\beta(1-\theta)/2}. \end{aligned}$$

Solving $1 + 2\theta - \beta(1-\theta)/2 < -1$ for θ, assuming $\beta > 0$ and $\theta > 0$, leads to $\theta \in (0, (\beta-4)/(\beta+4))$ with $\beta > 4$. Summability may be verified via the integral test. For the test define $c = 1 + 2\theta - \beta(1-\theta)/2$, then

$$\int_e^\infty \ln(y)\left(\frac{y}{\ln(y)}\right)^c dy = \int_1^\infty \frac{e^{t(1+c)}}{t^{c-1}} dt < \infty,$$

because $c < -1$ and the last integral being the integral exponential function. Using $h_n \sim (\ln(n)/n)^\theta$ and (A.21) leads to

$$\sup_{x\in\mathbb{S}} \left|S_{n,j}(x) - \nu_j f_{X_i}(x)\right| = \mathcal{O}\left(\left(\frac{\ln(n)}{n}\right)^\theta\right) + \mathcal{O}\left(\left(\frac{\ln(n)}{n}\right)^{(1-\theta)/2}\right) \qquad \text{almost surely.}$$

For $\theta \in (0, 1/3)$ and $\ln(n)/n \to 0$ as $n \to \infty$, $\theta < (1-\theta)/2$ for which the first term dominates the second. If $\theta \in [1/3, 1)$ the second term dominates the first which proves the second statement of the lemma

□

A.3 Proof of Theorem 2.3.7

According to (2.8) and $x \in \mathbb{S}$ it follows that

$$\begin{aligned}
L_n'(x;\lambda_n) &= \frac{1}{nh}\sum_{i=1}^n \frac{(X_i - x)K_{h_n}(X_i - x)}{1 + \lambda_n(x)(X_i - x)K_{h_n}(X_i - x)} \\
&= \frac{1}{nh}\left|\sum_{i=1}^n \frac{-(X_i - x)K_{h_n}(X_i - x)}{1 + \lambda_n(x)(X_i - x)K_{h_n}(X_i - x)}\right| \\
&= \frac{1}{nh}\left|\sum_{i=1}^n \left(\frac{\lambda_n(x)((X_i - x)K_{h_n}(X_i - x))^2}{1 + \lambda_n(x)(X_i - x)K_{h_n}(X_i - x)} - (X_i - x)K_{h_n}(X_i - x)\right)\right| \\
&\geq \frac{1}{nh}\left|\sum_{i=1}^n \frac{\lambda_n(x)((X_i - x)K_{h_n}(X_i - x))^2}{1 + \lambda_n(x)(X_i - x)K_{h_n}(X_i - x)}\right| - \left|S_{n,1}(x)\right| \\
&\geq \frac{|\lambda_n(x)|S_{n,2}(x)}{1 + C_4|\lambda_n(x)|} - \left|S_{n,1}(x)\right|,
\end{aligned}$$

with $C_4 = \sup_{u\in\mathbb{R}} uK(u)$ and $S_{n,j}(x)$ given in (2.11). The last line follows from $1 + \lambda_n(x)(X_i - x)K_h(X_i - x) = |1 + \lambda_n(x)(X_i - x)K_h(X_i - x)|$ since $p_i(x;\lambda_n) > 0$ and therefore $|1+\lambda_n(x)(X_i-x)K_h(X_i-x)| \leq 1+C_4|\lambda_n(x)|$. The above derivation is similar to Chen and Hall (1993, pp. 1174–1175), see also the proof of theorem 1 in Steikert (2014b). Since $L_n'(x;\lambda_n) = 0$,

$$|\lambda_n(x)| \leq \frac{|S_{n,1}(x)|}{S_{n,2}(x) - C_4|S_{n,1}(x)|},$$

the problem is therefore to show that

$$\sup_{x\in\mathbb{S}}|\lambda_n(x)| \leq \sup_{x\in\mathbb{S}}\left\{\frac{|S_{n,1}(x)|}{S_{n,2}(x)-C_4|S_{n,1}(x)|}\right\} =: A_n, \tag{A.22}$$

converges to zero almost surely. Note that

$$\begin{aligned} A_n &\leq \sup_{x\in\mathbb{S}}\big\{\big|S_{n,1}(x)\big|\big\}\sup_{x\in\mathbb{S}}\left\{\frac{1}{S_{n,2}(x)-C_4|S_{n,1}(x)|}\right\} \\ &=: A_{n,1}A_{n,2}. \end{aligned} \tag{A.23}$$

For $A_{n,1}$, it is easy to see that

$$\begin{aligned} A_{n,1} &= \sup_{x\in\mathbb{S}}\big\{\big|S_{n,1}(x)\big|\big\} \\ &= \sup_{x\in\mathbb{S}}\big\{\big|S_{n,1}(x)-\nu_1 f_{X_i}(x)\big|\big\} \\ &= \mathcal{O}(h_n)+\mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely,} \end{aligned} \tag{A.24}$$

with $\nu_1 = 0$ due to symmetry of the kernel function and the last line following from Lemma 2.3.6. To bound $A_{n,2}$ define

$$\Upsilon_n = \sup_{x\in\mathbb{S}}\Big|S_{n,2}(x)-\nu_2 f_{X_i}(x)-C_4\big|S_{n,1}(x)-\nu_1 f_{X_i}(x)\big|\Big|.$$

Then, due to Lemma 2.3.6, it is easy to see that

$$\begin{aligned} \Upsilon_n &\leq \sup_{x\in\mathbb{S}}\big\{\big|S_{n,2}(x)-\nu_2 f_{X_i}(x)\big|\big\} + C_4\sup_{x\in\mathbb{S}}\big\{\big|S_{n,1}(x)-\nu_1 f_{X_i}(x)\big|\big\} \\ &= \mathcal{O}(h_n)+\mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \end{aligned}$$

Thus, for $x\in\mathbb{S}$ and n sufficiently large,

$$\begin{aligned} -\Upsilon_n \leq S_{n,2}(x)-\nu_2 f_{X_i}(x) &- C_4\big|S_{n,1}(x)-\nu_1 f_{X_i}(x)\big| \leq \Upsilon_n \\ \Longleftrightarrow \nu_2 f_{X_i}(x)-\Upsilon_n \leq S_{n,2}(x) &- C_4\big|S_{n,1}(x)\big| \leq \nu_2 f_{X_i}(x)+\Upsilon_n, \end{aligned}$$

employing $\nu_1 = 0$. Because $S_{n,2}(x)-C_4|S_{n,1}(x)|>0$ and $f_{X_i}(x)>0$ on $\mathbb{S}$,

$$\frac{1}{S_{n,2}(x)-C_4\big|S_{n,1}(x)\big|} \leq \frac{1}{\nu_2 f_{X_i}(x)-\Upsilon_n}.$$

Combining this expression with $A_{n,2}$, defined in (A.23), it follows that for n sufficiently large,

$$\begin{aligned}
A_{n,2} &= \sup_{x\in\mathbb{S}}\left\{\frac{1}{S_{n,2}(x) - C_4|S_{n,1}(x)|}\right\} \\
&\leq \sup_{x\in\mathbb{S}}\left\{\frac{1}{\nu_2 f_{X_i}(x) - \Upsilon_n}\right\} \\
&\leq \frac{1}{C_8\nu_2 - \Upsilon_n} \\
&= \frac{1}{C_8\nu_2\Big(1 + \mathcal{O}(h_n) + \mathcal{O}\Big(\sqrt{\ln(n))/(nh_n)}\Big)\Big)} \quad \text{almost surely} \\
&= \frac{1}{C_8\nu_2}\left(1 + \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right)\right) \quad \text{almost surely,} \qquad \text{(A.25)}
\end{aligned}$$

because $0 < C_8 = \inf_{x\in\mathbb{S}} f_{X_i}(x)$ according to Assumption 2.3.3.c). Inserting $A_{n,1}$ and $A_{n,2}$, given in (A.24) and (A.25), in (A.23) leads to

$$\sup_{x\in\mathbb{S}}\left\{\frac{|S_{n,1}(x)|}{S_{n,2}(x) - C_4|S_{n,1}(x)|}\right\} = \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely,}$$

which in combination with (A.22) results in

$$\sup_{x\in\mathbb{S}}|\lambda_n(x)| = \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \qquad \text{(A.26)}$$

This completes the proof for the first statement.

For the second statement of the theorem note that

$$\begin{aligned}
\sup_{x\in\mathbb{S}}\left|p_i(x;\lambda_n) - \frac{1}{n}\right| &= \sup_{x\in\mathbb{S}}\left|\frac{1}{n(1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x))} - \frac{1}{n}\right| \\
&= \sup_{x\in\mathbb{S}}\left|\frac{-\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}{n(1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x))}\right| \\
&= \sup_{x\in\mathbb{S}}\left\{\big|\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)\big|\left|\frac{1}{n(1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x))}\right|\right\} \\
&\leq \sup_{x\in\mathbb{S}}\left\{C_4\big|\lambda_n(x)\big|\left|\frac{1}{n(1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x))}\right|\right\} \\
&\leq \sup_{x\in\mathbb{S}}\big\{C_4\big|\lambda_n(x)\big|\big\} \\
&= \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.}
\end{aligned}$$

The fourth line follows from $C_4 = \sup_{u\in\mathbb{R}} uK(u)$, the fifth from $p_i(x;\lambda_n)$ being probabilities, i.e., $p_i(x;\lambda_n) = (n(1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)))^{-1} \leq 1$, and the last from (A.26). This completes the proof.

□

A.4 Proof of Proposition 2.3.8

The proof bares some resemblance to the proof of Lemma 2.3.6. Rewriting the expression of the proposition leads to

$$\begin{aligned}
\sup_{x\in\mathbb{S}}\left|J_3(x) - f_{X_i}(x)\right| &= \sup_{x\in\mathbb{S}}\left|\sum_{i=1}^n p_i(x;\lambda_n)K_{h_n}(X_i-x) - f_{X_i}(x)\right| \\
&= \sup_{x\in\mathbb{S}}\left|\sum_{i=1}^n p_i(x;\lambda_n)K_{h_n}(X_i-x)\right. \\
&\qquad\left. - \frac{1}{n}\sum_{i=1}^n K_{h_n}(X_i-x) + \frac{1}{n}\sum_{i=1}^n K_{h_n}(X_i-x) - f_{X_i}(x)\right| \\
&\leq \sup_{x\in\mathbb{S}}\left|\sum_{i=1}^n\left(p_i(x;\lambda_n) - \frac{1}{n}\right)K_{h_n}(X_i-x)\right| \\
&\qquad + \sup_{x\in\mathbb{S}}\left|\frac{1}{n}\sum_{i=1}^n K_{h_n}(X_i-x) - f_{X_i}(x)\right| \\
&=: A_n + B_n. \qquad \text{(A.27)}
\end{aligned}$$

The second line follows from adding and subtracting the ordinary kernel density estimator $n^{-1}\sum_{i=1}^n K_{h_n}(X_i-x)$. For A_n note that,

$$\begin{aligned}
A_n &= \sup_{x\in\mathbb{S}}\left|\sum_{i=1}^n\left(p_i(x;\lambda_n) - \frac{1}{n}\right)K_{h_n}(X_i-x)\right| \\
&= \sup_{x\in\mathbb{S}}\left|\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)} - 1\right)K_{h_n}(X_i-x)\right| \\
&= \sup_{x\in\mathbb{S}}\left|\frac{1}{n}\sum_{i=1}^n\left(-\frac{\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}\right)K_{h_n}(X_i-x)\right| \\
&\leq \sup_{x\in\mathbb{S}}\left\{\frac{1}{n}\sum_{i=1}^n\left|\frac{\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}\right|K_{h_n}(X_i-x)\right\} \\
&\leq \sup_{x\in\mathbb{S}}\left\{C_4\left|\lambda_n(x)\right|\right\}\sup_{x\in\mathbb{S}}\left\{\frac{1}{n}\sum_{i=1}^n\left|\frac{1}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}\right|K_{h_n}(X_i-x)\right\} \\
&=: A_{n,1}A_{n,2}, \qquad \text{(A.28)}
\end{aligned}$$

with the last line following from the fact that $\sup_{u\in\mathbb{R}} uK(u) = C_4$. From the first part of Theorem 2.3.7 it is apparent that

$$A_{n,1} = \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \tag{A.29}$$

To bound $A_{n,2}$ define

$$\Delta_n = \sup_{x\in\mathbb{S}} \left|\lambda_n(x)(X_i - x)K_{h_n}(X_i - x)\right|.$$

Then, because $\Delta_n \leq \sup_{x\in\mathbb{S}} |C_4\lambda_n(x)|$ and $\sup_{x\in\mathbb{S}} |\lambda_n(x)| = \mathcal{O}(h_n) + \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ almost surely, due to Theorem 2.3.7, $\Delta_n = \mathcal{O}(h_n) + \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ almost surely. Thus, for n sufficiently large and $x \in \mathbb{S}$, it follows that

$$1 - \Delta_n \leq 1 + \lambda_n(x)(X_i - x)K_{h_n}(X_i - x) \leq 1 + \Delta_n,$$

which implies

$$\frac{1}{1 + \lambda_n(x)(X_i - x)K_{h_n}(X_i - x)} \leq \frac{1}{1 - \Delta_n}, \tag{A.30}$$

since $1 + \lambda_n(x)(X_i - x)K_{h_n}(X_i - x) > 0$ because $p_i(x; \lambda_n) > 0$. For n sufficiently large,

$$\begin{aligned} A_{n,2} &= \sup_{x\in\mathbb{S}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{1 + \lambda_n(x)(X_i - x)K_{h_n}(X_i - x)} \right| K_{h_n}(X_i - x) \right\} \\ &\leq \frac{1}{1 - \Delta_n} \sup_{x\in\mathbb{S}} \left\{ \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(X_i - x) \right\} \\ &=: A_{n,21} A_{n,22}. \end{aligned} \tag{A.31}$$

To bound $A_{n,21}$ it is easy to see that

$$A_{n,21} = 1 + \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely,} \tag{A.32}$$

because $\Delta_n = \mathcal{O}(h_n) + \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ almost surely. To determine $A_{n,22}$ note that

$$\begin{aligned} A_{n,22} &= \sup_{x\in\mathbb{S}} \left\{ \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(X_i - x) \right\} \\ &= \sup_{x\in\mathbb{S}} \left\{ \frac{1}{n} \left| \sum_{i=1}^{n} (K_{h_n}(X_i - x) - \mathbb{E}\, K_{h_n}(X_1 - x) + \mathbb{E}\, K_{h_n}(X_1 - x)) \right| \right\} \end{aligned}$$

$$\leq \sup_{x\in\mathbb{S}}\left\{\frac{1}{n}\Big|\sum_{i=1}^{n}(K_{h_n}(X_i-x)-\mathbb{E}\,K_{h_n}(X_1-x))\Big|\right\}+\sup_{x\in\mathbb{S}}\{\mathbb{E}\,K_{h_n}(X_1-x)\}$$
$$=: A_{n,221}+A_{n,222}. \tag{A.33}$$

The second line follows from adding and subtracting $\mathbb{E}\,K_{h_n}(X_1-x)$ and the kernel function being a density.

To determine the first term of (A.33) similar steps as in the proof of Lemma 2.3.6 are followed. For this define $S_n(x)=n^{-1}\sum_{i=1}^{n}K_h(X_i-x)$ and note the similarity to (2.11). Again, cover $\mathbb{S}$ by a finite number, L_n, of subintervals I_k with centers x_k and length l_k, where $k=1,...,L_n$. Similar to (A.3), $A_{n,221}$ is bounded as follows

$$\begin{aligned}A_{n,221}\leq \max_{1\leq k\leq L_n}\left\{\sup_{x\in\mathbb{S}\cap I_k}\big|S_n(x)-S_n(x_k)\big|\right\}+\max_{1\leq k\leq L_n}\big|S_n(x_k)-\mathbb{E}\,S_n(x_k)\big|\\ +\max_{1\leq k\leq L_n}\sup_{x\in\mathbb{S}\cap I_k}\big|\,\mathbb{E}\,S_n(x_k)-\mathbb{E}\,S_n(x)\big|\\ =: A_{n,2211}+A_{n,2212}+A_{n,2213}.\end{aligned} \tag{A.34}$$

Following similar arguments that resulted in (A.4) and (A.5) lead to

$$A_{n,2211}=\mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right)\quad\text{almost surely}\quad\text{and}\quad A_{n,2213}=\mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right). \tag{A.35}$$

Regarding $A_{n,2212}$ the summability of

$$\mathbb{P}(A_{n,2212}>\varepsilon)\leq L_n\sup_{x\in\mathbb{R}}\mathbb{P}(|S_n(x)-\mathbb{E}\,S_n(x)|>\varepsilon),$$

which is similar to (A.6), is established below in order to apply the Borel-Cantelli lemma. To avoid redundancies in the application of Lemma B.1 only the necessary steps are provided here. Define $Z_i(x)=K((X_i-x)/h_n)-\mathbb{E}\,K((X_1-x)/h_n)$ similar as in (A.7), then, it is easy to see that, $|Z_i(x)|\leq 2C_1=:b$ using Assumption 2.3.2.a). To determine $\sigma^2_{s_n}$ follow similar steps leading to (A.14). First, as in (A.9), rewrite the variance as follows

$$\begin{aligned}\sigma^2_{s_n}&\leq s_n\,\mathbb{E}\,Z^2_{j,1}(x)+2s_n\sum_{i=2}^{s_n}\big|\,\mathbb{E}\,Z_{j,1}(x)Z_{j,i}(x)\big|\\ &=:\Sigma_1+\Sigma_2.\end{aligned}$$

Then for Σ_1, following similar steps leading to (A.10), result in

$$\Sigma_1\leq s_nh_n(C_6+\mathcal{O}(h_n)),$$

using $\int K^2(u)du \leq 1$. Then, for Σ_2,

$$\begin{aligned}\Sigma_2 &\leq 2s_n \sum_{i=2}^{\lfloor h_n^{-1} \rfloor} \left| \mathbb{E}\, Z_{j,1}(x) Z_{j,i}(x) \right| + 2s_n \sum_{i=\lfloor h_n^{-1} \rfloor + 1}^{\infty} \left| \mathbb{E}\, Z_{j,1}(x) Z_{j,i}(x) \right| \\ &=: \Sigma_{21} + \Sigma_{22}.\end{aligned}$$

For Σ_{21}, similar to (A.12), it follows that

$$\Sigma_{21} \leq \text{constant}_1 \times s_n h_n,$$

using $\int K(u)du = 1$. Similar to (A.13),

$$\Sigma_{22} \leq \text{constant}_2 \times s_n h_n \sum_{k=\lfloor h_n^{-1} \rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta},$$

because $\mathbb{E}\,|K((X_1 - x)/h_n)| \leq C_6 h_n$ and an application of Davydov's lemma. Thus,

$$\sigma_{s_n}^2 \leq s_n h_n \left(C_6 + \text{constant}_1 + \text{constant}_2 \times \sum_{k=\lfloor h_n^{-1} \rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta} + \mathcal{O}(h_n) \right).$$

Define $\varepsilon = \varepsilon_n = C_\varepsilon \sqrt{\ln(n)/(nh_n)}$ and $s_n = \lfloor \sqrt{nh_n / \ln(n)} \rfloor$ which need to be as in the proof of Lemma (2.3.6). By applying Lemma B.1 in Appendix B it follows that

$$A_{n,2212} = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \quad \text{almost surely,} \tag{A.36}$$

for sufficiently large C_ε. Combining (A.34) with $A_{n,2211}$ and $A_{n,2213}$, both given in (A.35), and $A_{n,2212}$, defined in (A.36), leads to

$$A_{n,221} = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \qquad \text{almost surely.} \tag{A.37}$$

To determine the second term of (A.33), i.e., $A_{n,222}$, note that

$$\begin{aligned}A_{n,222} &= \sup_{x \in \mathbb{S}} \left\{ \frac{1}{h_n} \int K\left(\frac{z - x}{h_n} \right) f_{X_i}(z) dz \right\} \\ &\leq \sup_{x \in \mathbb{S}} \left\{ \int K(u) f_{X_i}(x + uh_n) du \right\} \\ &\leq C_6. \end{aligned} \tag{A.38}$$

The second line follows from a change of variable, i.e., $z = x + uh_n$. The last line is due to the marginal density being bounded by C_6 and the fact that $\int K(u)du = 1$. Substituting (A.37) and (A.38) into (A.33) leads to

$$A_{n,22} \leq C_6 + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \tag{A.39}$$

Plugging (A.32) and (A.39) into (A.31) proves

$$A_{n,2} \leq C_6\left(1 + \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right)\right) \qquad \text{almost surely.} \tag{A.40}$$

Finally, to determine A_n, given in (A.28), use (A.29) and (A.40), then

$$A_n = \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \tag{A.41}$$

Regarding B_n, the second term of (A.27), note that

$$\begin{aligned}
B_n &= \sup_{x\in\mathbb{S}}\left|\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(X_i - x) - f_{X_i}(x)\right| \\
&= \sup_{x\in\mathbb{S}}\left|\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(X_i - x) - \mathbb{E}\,K_{h_n}(X_1 - x) + \mathbb{E}\,K_{h_n}(X_1 - x) - f_{X_i}(x)\right| \\
&\leq \sup_{x\in\mathbb{S}}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(X_i - x) - \mathbb{E}\,K_{h_n}(X_1 - x)\right|\right\} + \sup_{x\in\mathbb{S}}\left|\,\mathbb{E}\,K_{h_n}(X_1 - x) - f_{X_i}(x)\right| \\
&= B_{n,1} + B_{n,2}.
\end{aligned} \tag{A.42}$$

The first term, $B_{n,1}$, is equivalent to $A_{n,221}$, given in (A.33), thus, according to (A.37),

$$B_{n,1} = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \tag{A.43}$$

For $B_{n,2}$ it is easy to see that

$$\begin{aligned}
B_{n,2} &= \sup_{x\in\mathbb{S}}\left|\,\mathbb{E}\,K_{h_n}(X_1 - x) - f_{X_i}(x)\right| \\
&\leq h_n C_7 \int |u| K(u)du \\
&= \mathcal{O}(h_n),
\end{aligned} \tag{A.44}$$

using Assumption 2.3.3.b). Combining (A.42), (A.43), and (A.44) leads to

$$B_n = \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \tag{A.45}$$

Finally, using (A.27), (A.41), and (A.45) proves the proposition, i.e.,

$$\sup_{x\in\mathbb{S}} \left|J_3(x) - f_{X_i}(x)\right| = \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.}$$

□

A.5 Proof of Corollary 2.3.9

Define

$$\Psi_n = \sup_{x\in\mathbb{S}} \left|J_3(x) - f_{X_i}(x)\right|,$$

then $\Psi_n = \mathcal{O}(h_n) + \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ with probability one due to Proposition 2.3.8. Note that $J_3(x) > 0$ because existence of the estimator $\widehat{m}(x)$ is presumed, $p_i(x;\lambda_n) > 0$, and $K_{h_n}(X_i - x) \geq 0$. For n sufficiently large, it is easy to see that, $J_3^{-1}(x) \leq (f_{X_i}(x) - \Psi_n)^{-1}$ and therefore

$$\begin{aligned}
\sup_{x\in\mathbb{S}} \left|\frac{1}{J_3(x)}\right| &\leq \sup_{x\in\mathbb{S}} \left\{\frac{1}{f_{X_i}(x) - \Psi_n}\right\} \\
&\leq \frac{1}{C_8 - \Psi_n} \\
&\leq \frac{1}{C_8}\left(1 + \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right)\right) \qquad \text{almost surely,}
\end{aligned}$$

with C_8 given in Assumption 2.3.3.c). This completes the proof.

□

A.6 Proof of Proposition 2.3.10

The proof is similar to the proof of the previous proposition and therefore I only provide the necessary steps. The notation, however, defining similar expressions as in Proposition 2.3.8 for which the proofs are compatible, such as $A_{n,1}$ or $B_{n,2}$, is kept for the present proof. For more details consult the equivalent expression in the proof of Proposition 2.3.8.

The expression of the proposition is rewritten in the following way

$$\begin{aligned}
&\sup_{x\in\mathbb{S}}\left|J_2(x)-\frac{h_n^2}{2}\mu_1 m''(x)f_{X_i}(x)\right| \\
&\qquad=\sup_{x\in\mathbb{S}}\left|\sum_{i=1}^{n}(m(X_i)-m(x))p_i(x;\lambda_n)K_{h_n}(X_i-x)-\frac{h_n^2}{2}\mu_1 m''(x)f_{X_i}(x)\right| \\
&\qquad=\sup_{x\in\mathbb{S}}\left\{\left|\frac{m''(x)}{2}\sum_{i=1}^{n}(X_i-x)^2 p_i(x;\lambda_n)K_{h_n}(X_i-x.\right.\right. \\
&\qquad\qquad\left.\left.-\frac{h_n^2}{2}\mu_1 m''(x)f_{X_i}(x)+o(h_n^2)\right|\right\} \\
&\qquad\le\sup_{x\in\mathbb{S}}\left|\frac{m''(x)}{2}\sum_{i=1}^{n}\left(p_i(x;\lambda_n)-\frac{1}{n}\right)(X_i-x)^2K_{h_n}(X_i-x)\right| \\
&\qquad\quad+\sup_{x\in\mathbb{S}}\left|\frac{m''(x)}{2n}\sum_{i=1}^{n}(X_i-x)^2K_{h_n}(X_i-x)-\frac{h_n^2}{2}\mu_1 m''(x)f_{X_i}(x)\right|+o(h_n^2) \\
&\qquad=:A_n+B_n+o(h_n^2). \qquad\text{(A.46)}
\end{aligned}$$

The second equation follows from a Taylor approximation of $m(X_i)$ possible due to Assumption 2.3.1.c) and condition (2.6). For the first inequality add and subtract $m''(x)(2n)^{-1}\sum_{i=1}^{n}(X_i-x)^2K_{h_n}(X_i-x)$.

Treating the terms A_n and B_n separately it follows that

$$\begin{aligned}
A_n&=\sup_{x\in\mathbb{S}}\left|\frac{m''(x)}{2n}\sum_{i=1}^{n}\left(\frac{-\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}\right)(X_i-x)^2K_{h_n}(X_i-x)\right| \\
&\le h_n^2\sup_{x\in\mathbb{S}}\left\{\frac{C_4\left|m''(x)\lambda_n(x)\right|}{2}\right\} \\
&\quad\times\sup_{x\in\mathbb{S}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}\right|\left(\frac{X_i-x}{h_n}\right)^2K_{h_n}(X_i-x)\right\} \\
&=:h_n^2A_{n,1}A_{n,2}. \qquad\text{(A.47)}
\end{aligned}$$

To bound $A_{n,1}$ note that

$$\begin{aligned}
A_{n,1}&=\sup_{x\in\mathbb{S}}\left\{\frac{C_4\left|m''(x)\lambda_n(x)\right|}{2}\right\} \\
&=\overline{C}\sup_{x\in\mathbb{S}}\left\{\left|\lambda_n(x)\right|\right\} \\
&=\mathcal{O}(h_n)+\mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right)\quad\text{almost surely,} \qquad\text{(A.48)}
\end{aligned}$$

because $|m''(x)|$ is bounded on $\mathbb{R}$ due Assumption 2.3.1.c) and Theorem 2.3.7. To bound $A_{n,2}$, define

$$\Delta_n = \sup_{x\in\mathbb{S}} \left|\lambda_n(x)(X_i - x)K_{h_n}(X_i - x)\right|,$$

then, due to (A.26), $\Delta_n = \mathcal{O}(h_n) + \mathcal{O}(\sqrt{\ln(n)/(nh_n)})$ almost surely. For sufficiently large n and the inequality given in (A.30), it follows that

$$\begin{aligned} A_{n,2} &\leq \frac{1}{1-\Delta_n} \sup_{x\in\mathbb{S}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\frac{X_i - x}{h_n} \right)^2 K_{h_n}(X_i - x) \right\} \\ &=: A_{n,21} A_{n,22}. \end{aligned} \tag{A.49}$$

The first term, $A_{n,21}$, is equivalent to (A.32) and therefore

$$A_{n,21} = 1 + \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \quad \text{almost surely.} \tag{A.50}$$

For the second term of (A.49) expand the expression by $\mathbb{E}((X_1 - x)/h_n)^2 K_{h_n}(X_1 - x)$ and note the similarities to (A.33). Thus,

$$\begin{aligned} A_{n,22} &= \sup_{x\in\mathbb{S}} \left\{ \frac{1}{n} \left| \sum_{i=1}^{n} \left(\left(\frac{X_i - x}{h_n} \right)^2 K_{h_n}(X_i - x) - \mathbb{E}\left(\frac{X_1 - x}{h_n} \right)^2 K_{h_n}(X_1 - x) \right.\right.\right. \\ &\qquad \left.\left.\left. + \mathbb{E}\left(\frac{X_1 - x}{h_n} \right)^2 K_{h_n}(X_1 - x) \right| \right) \right\} \\ &\leq \sup_{x\in\mathbb{S}} \left\{ \frac{1}{nh_n} \left| \sum_{i=1}^{n} Z_i(x) \right| \right\} + \sup_{x\in\mathbb{S}} \left\{ \mathbb{E}\left(\frac{X_1 - x}{h_n} \right)^2 K_{h_n}(X_1 - x) \right\} \\ &=: A_{n,221} + A_{n,222}, \end{aligned} \tag{A.51}$$

with $Z_i(x) = ((X_i - x)/h_n)^2 K((X_i - x)/h_n) - \mathbb{E}((X_1 - x)/h_n)^2 K((X_1 - x)/h_n)$. Note that $Z_i(x) \leq Z_{1,i}(x)$, with $Z_{1,i}(x)$ defined in (A.7), because $(X_i - x)/h_n$ is bounded by 1 due to Assumption 2.3.2.b). Thus, the result of $Z_{1,i}(x)$ translates to $Z_i(x)$ implying

$$A_{n,221} = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \quad \text{almost surely,} \tag{A.52}$$

using condition i) of Lemma 2.3.6. To determine $A_{n,222}$ of (A.51) similar steps leading to

(A.38) are used to find

$$A_{n,222} \leq C_6 \mu_1, \tag{A.53}$$

with $\mu_1 = \int u^2 K(u) du$. Combining (A.51), (A.52), and (A.53) result in

$$A_{n,22} \leq C_6 \mu_1 + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{n h_n}}\right) \quad \text{almost surely.} \tag{A.54}$$

Substituting (A.50) and (A.54) in (A.49) lead to

$$A_{n,2} \leq C_6 \mu_1 \left(1 + \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{n h_n}}\right)\right) \quad \text{almost surely.} \tag{A.55}$$

Finally, using (A.47), (A.48), and (A.55) to determine A_n it follows that

$$A_n = \mathrm{o}(h_n^2) \quad \text{almost surely.} \tag{A.56}$$

For B_n, the second summand of (A.46), note that

$$\begin{aligned} B_n &= \sup_{x \in \mathbb{S}} \left| \frac{m''(x)}{2n} \sum_{i=1}^{n} (X_i - x)^2 K_{h_n}(X_i - x) - \frac{h_n^2}{2} \mu_1 m''(x) f_{X_i}(x) \right| \\ &\leq \sup_{x \in \mathbb{S}} \left| \frac{m''(x)}{2n} \sum_{i=1}^{n} \Big((X_i - x)^2 K_{h_n}(X_i - x) - \mathbb{E}(X_1 - x)^2 K_{h_n}(X_1 - x) \Big) \right| \\ &\quad + \sup_{x \in \mathbb{S}} \left| \frac{m''(x)}{2n} \sum_{i=1}^{n} \mathbb{E}(X_1 - x)^2 K_{h_n}(X_1 - x) - \frac{h_n^2}{2} \mu_1 m''(x) f_{X_i}(x) \right| \\ &=: B_{n,1} + B_{n,2}. \end{aligned} \tag{A.57}$$

For $B_{n,1}$ define $Z_i(x)$ as before and note the similarities to $A_{n,221}$ in (A.51), then

$$\begin{aligned} B_{n,1} &= \sup_{x \in \mathbb{S}} \left| \frac{h_n^2 m''(x)}{2n} \sum_{i=1}^{n} \left(\left(\frac{X_i - x}{h_n} \right)^2 K_{h_n}(X_i - x) - \mathbb{E}\left(\frac{X_1 - x}{h_n} \right)^2 K_{h_n}(X_1 - x) \right) \right| \\ &= \frac{h_n^2}{2} \sup_{x \in \mathbb{S}} \left| \frac{m''(x)}{n h_n} \sum_{i=1}^{n} Z_i(x) \right| \\ &= h_n^2 \, \mathcal{O}\left(\sqrt{\frac{\ln(n)}{n h_n}} \right) = \mathrm{o}(h_n^2) \qquad \text{almost surely.} \end{aligned} \tag{A.58}$$

The last line follows from Assumption 2.3.1.c) and (A.52). To determine $B_{n,2}$ of (A.57) it

is easy to see that

$$\begin{aligned} B_{n,2} &= \sup_{x\in\mathbb{S}} \left| \frac{h_n^2 m''(x)}{2} \left(\mathbb{E}\left(\frac{X_1 - x}{h_n} \right)^2 K_{h_n}(X_1 - x) - \mu_1 f_{X_i}(x) \right) \right| \\ &= \sup_{x\in\mathbb{S}} \left| \frac{h_n^2 m''(x)}{2} \left(\int u^2 K(u) f_{X_i}(x + uh_n) - f_{X_i}(x) \right) du \right| \\ &\leq \overline{C} h_n^3 \int |u|^3 K(u) du \\ &= \mathcal{O}(h_n^3) = \mathrm{o}(h_n^2), \end{aligned} \tag{A.59}$$

with $\overline{C}$ denoting a suitable constant. The second line follows from a change of variable, i.e., $u = (z-x)/h_n$, the third because $|m''(x)|$ is bounded on $\mathbb{R}$ and the marginal density is Lipschitz continuous. Thus, for B_n, given in (A.57), it follows, using the results of (A.58) and (A.59), that

$$B_n = \mathrm{o}(h_n^2) \qquad \text{almost surely.} \tag{A.60}$$

Combining (A.46), (A.56), and (A.60) lead to

$$\sup_{x\in\mathbb{S}} \left| J_2(x) - \frac{h_n^2}{2} \mu_1 m''(x) f_{X_i}(x) \right| = \mathrm{o}(h_n^2) \qquad \text{almost surely,}$$

which completes the proof.

□

A.7 Proof of Corollary 2.3.11

The proof is similar to the proof of the previous corollary. Define

$$\Phi_n = \sup_{x\in\mathbb{S}} \left| J_2(x) - \frac{h_n^2}{2} \mu_1 m''(x) f_{X_i}(x) \right|,$$

then, given Proposition 2.3.10, $\Phi_n = \mathrm{o}(h_n^2)$ almost surely. For n sufficiently large, it is easy to see that,

$$J_2(x) \leq \frac{h_n^2}{2} \mu_1 m''(x) f_{X_i}(x) + \Phi_n,$$

and therefore

$$\sup_{x\in\mathbb{S}} \left| J_2(x) \right| \leq \overline{C} h_n^2 + \Phi_n = \mathcal{O}(h_n^2) \qquad \text{almost surely,}$$

with suitable constant $\overline{C}$. Note that $f_{X_i}(x)$ and $m''(x)$ are bounded on $\mathbb{S}$ due to Assumptions 2.3.3.a) and 2.3.1.c). This completes the proof.

□

A.8 Proof of Lemma 2.3.12

The proof is the same as the proof of Lemma 3 in Steikert (2014b). It is repeated for the readers convenience. Proving $|\phi(X_{i+1})| \leq \tau_n$ almost surely for $i \leq n$ and n sufficiently large using the Borel-Cantelli lemma suffices to prove the statement. Note that

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(|\phi(X_{n+1})| > \tau_n) &\leq \sum_{n=1}^{\infty} \frac{\mathbb{E}\,|\phi(X_{n+1})|^s}{\tau_n^s} \\ &= \mathbb{E}\,|\phi(X_2)|^s \sum_{n=1}^{\infty} \tau_n^{-s} < \infty,\end{aligned}$$

where the first line follows from Markov's inequality and the second from the finiteness of $\mathbb{E}\,|\phi(X_2)|^s$ for some $s > 2$, stationarity, and the summability of τ_n^{-s}, with τ_n given in (2.16). The Summability of τ_n^{-s} is established using the integral test. For $n > e$,

$$\int_n^{\infty} \frac{1}{y(\ln(\ln(y)))^2 \ln(y)} dy = -\frac{1}{\ln(\ln(y))}\Big|_n^{\infty} = \frac{1}{\ln(\ln(n))} < \infty,$$

and therefore τ_n^{-s} is summable. Thus, $|\phi(X_{n+1})| \leq \tau_n$ with probability one for sufficiently large n. Because τ_n is increasing in n, $|\phi(X_{i+1})| \leq \tau_n$ with probability one for $i \leq n$ and n sufficiently large. This completes the proof.

□

A.9 Proof of Corollary 2.3.13

The corollary is a straightforward implication of Lemma 2.3.12. Because $|\phi(X_{i+1})| \leq \tau_n$ almost surely for $i \leq n$ and n sufficiently large the difference $T_n(x) - T_n^{(t)}(x)$ is eventually zero with probability one.

□

A.10 Proof of Lemma 2.3.14

The proof is similar to the proof of Lemma 2.3.6, therefore I only provide the necessary steps. Because $\mathbb{S}$ is a compact real interval it may be covered by a finite number, Λ_n, of subintervals I_k with centers x_k and length l_n, where $k = 1, ..., \Lambda_n$. Then,

$$\begin{aligned}
\sup_{x\in\mathbb{S}} \left| T_n^{(t)}(x) \right| &= \max_{1\le k\le\Lambda_n} \sup_{x\in\mathbb{S}\cap I_k} \left| T_n^{(t)}(x) \right| \\
&= \max_{1\le k\le\Lambda_n} \sup_{x\in\mathbb{S}\cap I_k} \left| T_n^{(t)}(x) - T_n^{(t)}(x_k) + T_n^{(t)}(x_k) \right| \\
&\le \max_{1\le k\le\Lambda_n} \left\{ \sup_{x\in\mathbb{S}\cap I_k} \left| T_n^{(t)}(x) - T_n^{(t)}(x_k) \right| \right\} + \max_{1\le k\le\Lambda_n} \left| T_n^{(t)}(x_k) \right| \\
&=: A_n + B_n. \qquad \text{(A.61)}
\end{aligned}$$

For A_n, using the definitions (2.17) and (2.18), it is easy to see that,

$$\begin{aligned}
A_n &= \max_{1\le k\le\Lambda_n} \sup_{x\in\mathbb{S}\cap I_k} \left| \frac{1}{nh_n} \sum_{i=1}^{n} \Big(\phi(X_{i+1}) \mathbb{1}_{\{|\phi(X_{i+1})|\le\tau_n\}} - m^{(t)}(X_i) \Big) \right. \\
&\qquad \left. \times \left(K\left(\frac{X_i - x}{h_n} \right) - K\left(\frac{X_i - x_k}{h_n} \right) \right) \right| \\
&\le \max_{1\le k\le\Lambda_n} \sup_{x\in\mathbb{S}\cap I_k} \left\{ \frac{1}{nh_n} \sum_{i=1}^{n} \Big(\left| \phi(X_{i+1}) \mathbb{1}_{\{|\phi(X_{i+1})|\le\tau_n\}} \right| + \left| m^{(t)}(X_i) \right| \Big) \right. \\
&\qquad \left. \times \left| K\left(\frac{X_i - x}{h_n} \right) - K\left(\frac{X_i - x_k}{h_n} \right) \right| \right\} \\
&\le \max_{1\le k\le\Lambda_n} \sup_{x\in\mathbb{S}\cap I_k} \left\{ \frac{C_2 |x_k - x|}{nh_n^2} \sum_{i=1}^{n} \Big(\left| \phi(X_{i+1}) \mathbb{1}_{\{|\phi(X_{i+1})|\le\tau_n\}} \right| + \left| m^{(t)}(X_i) \right| \Big) \right\} \\
&\le \max_{1\le k\le\Lambda_n} \sup_{x\in\mathbb{S}\cap I_k} \left\{ \frac{2C_2\tau_n}{h_n^2} |x_k - x| \right\} \\
&\le \frac{2C_2\tau_n l_n}{h_n^2} \\
&= \frac{\overline{C}\tau_n}{h_n^2\Lambda_n} \\
&= \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \qquad \text{almost surely,} \qquad \text{(A.62)}
\end{aligned}$$

with suitable constant $\overline{C}$. The second inequality follows from Assumption 2.3.2.c) and the third because $\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\le\tau_n\}}$ and $m^{(t)}(X_i)$ are both, by definition, bounded by τ_n. The second to last line follows from the definition of the length l_n, i.e., $l_n = \overline{C}/\Lambda_n$ with suitable constant $\overline{C}$. For the last line define $\Lambda_n = \lceil \tau_n \sqrt{n/(h_n^3 \ln(n))} \rceil$ and note that

$\tau_n\sqrt{n/(h_n^3\ln(n))} \leq \Lambda_n$.

For B_n, the second summand of (A.61), proving the summability of

$$\begin{aligned}\mathbb{P}(B_n > \varepsilon) &= \mathbb{P}\Big(\max_{1\leq k\leq\Lambda_n} \big|\, T_n^{(t)}(x_k)\big| > \varepsilon\Big) \\ &\leq \sum_{k=1}^{\Lambda_n} \mathbb{P}\Big(\big|\, T_n^{(t)}(x_k)\big| > \varepsilon\Big) \\ &\leq \Lambda_n \sup_{x\in\mathbb{S}} \mathbb{P}\Big(\big|\, T_n^{(t)}(x)\big| > \varepsilon\Big) \\ &= \Lambda_n \sup_{x\in\mathbb{S}} \left\{ \mathbb{P}\left(\left| \sum_{i=1}^{n} U_i(x) \right| > \varepsilon n h_n \right) \right\}, \end{aligned} \tag{A.63}$$

with

$$U_i(x) = \Big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}} - m^{(t)}(X_i)\Big) K\left(\frac{X_i - x}{h_n}\right),$$

and appropriate ε is necessary. Then, by virtue of the Borel-Cantelli lemma the statement of the lemma follows.

To bound the probability term of (A.63) I again employ Lemma B.1 in Appendix B. For this I specify the bound b, the variance $\sigma_{r_n}^2$, and $r_n \leq n$ such that $r_n < \varepsilon n h_n b/4$ in the following way. To determine b note that

$$\begin{aligned}\big|U_i(x)\big| &\leq \big|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}} - m^{(t)}(X_i)\big| K\left(\frac{X_i - x}{h_n}\right) \\ &\leq \Big(\big|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\big| + \big|m^{(t)}(X_i)\big|\Big) K\left(\frac{X_i - x}{h_n}\right) \\ &\leq 2\tau_n K\left(\frac{X_i - x}{h_n}\right) \\ &\leq 2C_1\tau_n =: b, \end{aligned}$$

using Assumption 2.3.2.a) and the fact that $\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}$ and $m^{(t)}(X_i)$ are both, by definition, bounded by τ_n.

For the variance $\sigma_{r_n}^2$ note that, similar to (A.9),

$$\begin{aligned}\sigma_{r_n}^2 = \mathbb{E}\left(\sum_{i=1}^{r_n} U_i(x)\right)^2 &\leq r_n \mathrm{var}(U_1(x)) + 2r_n \sum_{i=2}^{r_n} \big|\mathrm{cov}(U_1(x), U_i(x))\big| \\ &=: \Sigma_1 + \Sigma_2. \end{aligned} \tag{A.64}$$

To bound Σ_1 note that

$$\begin{aligned}
\Sigma_1 &= r_n \mathrm{var}\bigg(\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}} - m^{(t)}(X_1)\Big)K\bigg(\frac{X_1-x}{h_n}\bigg)\bigg) \\
&= r_n\,\mathbb{E}\bigg(\Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}} - m^{(t)}(X_1)\Big)K\bigg(\frac{X_1-x}{h_n}\bigg)\bigg)^2 \\
&\leq r_n\,\mathbb{E}\Big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big| + \big|m^{(t)}(X_1)\big|\Big)^2 K^2\bigg(\frac{X_1-x}{h_n}\bigg) \\
&\leq r_n\,\mathbb{E}\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big| + C_{11}\big)^2 K^2\bigg(\frac{X_1-x}{h_n}\bigg) \\
&\leq r_n\,\mathbb{E}\bigg(\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big| + C_{11}\big)^2\,\mathbb{E}\bigg(K^2\bigg(\frac{X_1-x}{h_n}\bigg)\bigg|X_2\bigg)\bigg) \\
&\leq r_n\,\mathbb{E}\bigg(\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big| + C_{11}\big)^2 \int K^2\bigg(\frac{z-x}{h_n}\bigg) f_{X_1|X_2}(z|X_2)dz\bigg) \\
&\leq C_{12} r_n h_n\,\mathbb{E}\big(\big|\phi(X_2)\big| + C_{11}\big)^2 \int K^2(u)du \\
&= \text{constant}_1 \times r_n h_n. \qquad \text{(A.65)}
\end{aligned}$$

The fourth line follows from $\sup_{x\in\mathbb{S}}\{\sup_{\{u\in\mathbb{R}:|u-x|\leq b_k h_n\}}|m^{(t)}(u)|\} = C_{11} < \infty$ due to Assumptions 2.3.2.a) and 2.3.2.b). The last line follows from the usual change of variable, $\mathbb{E}\,|\phi(X_{i+1})|^s$ for $s > 2$, and the fact that Assumptions 2.3.1.a), 2.3.3.a), 2.3.3.c), and 2.3.3.d) imply a bounded conditional density, i.e., $f_{X_i|X_{i+1}}(\cdot|v) = f_{X_i,X_{i+1}}(\cdot,v)/f_{X_{i+1}}(v) \leq C_{12}$.

To determine Σ_2 of (A.64) note that

$$\begin{aligned}
\Sigma_2 &\leq 2r_n \sum_{i=2}^{\lfloor h_n^{-1}\rfloor} \big|\mathrm{cov}(U_1(x), U_i(x))\big| + 2r_n \sum_{i=\lfloor h_n^{-1}\rfloor+1}^{\infty} \big|\mathrm{cov}(U_1(x), U_i(x))\big| \\
&=: \Sigma_{21} + \Sigma_{22}. \qquad \text{(A.66)}
\end{aligned}$$

Following similar arguments that resulted in (A.65) lead to

$$\begin{aligned}
\Sigma_{21} &\leq 2r_n \sum_{i=2}^{\lfloor h_n^{-1}\rfloor} \mathbb{E}\,\bigg(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}} - m^{(t)}(X_1)\big|\big|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}} - m^{(t)}(X_i)\big| \\
&\qquad \times K\bigg(\frac{X_1-x}{h_n}\bigg)K\bigg(\frac{X_i-x}{h_n}\bigg)\bigg) \\
&\leq 2r_n \sum_{i=2}^{\lfloor h_n^{-1}\rfloor} \mathbb{E}\,\bigg(\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big| + C_{11}\big)\big(\big|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\big| + C_{11}\big) \\
&\qquad \times\,\mathbb{E}\bigg(K\bigg(\frac{X_1-x}{h_n}\bigg)K\bigg(\frac{X_i-x}{h_n}\bigg)\bigg|X_2, X_{i+1}\bigg)\bigg)
\end{aligned}$$

$$
\begin{aligned}
&= 2r_n \sum_{i=2}^{\lfloor h_n^{-1} \rfloor} \mathbb{E}\left(\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big| + C_{11}\big)\big(\big|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\big| + C_{11}\big) \right. \\
&\qquad\qquad \left. \times \int K\Big(\frac{z-x}{h_n}\Big) K\Big(\frac{w-x}{h_n}\Big) f_{X_1,X_i|X_2,X_{i+1}}(z,w)dzdw \right) \\
&\leq 2C_{10} r_n h_n^2 \sum_{i=2}^{\lfloor h_n^{-1} \rfloor} \mathbb{E}\big(\big|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}}\big| + C_{11}\big)\big(\big|\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}\big| + C_{11}\big) \\
&\qquad\qquad \times \left(\int K(u)du \right)^2 \\
&\leq 2C_{10} r_n h_n \, \mathbb{E}\big(\big|\phi(X_2)\big| + C_{11}\big)\big(\big|\phi(X_{i+1})\big| + C_{11}\big) \\
&\leq 2C_{10} r_n h_n \sqrt{\mathbb{E}\big(\big|\phi(X_2)\big| + C_{11}\big)^2 \, \mathbb{E}\big(\big|\phi(X_{i+1})\big| + C_{11}\big)^2} \\
&= \text{constant}_2 \times r_n h_n. \qquad \text{(A.67)}
\end{aligned}
$$

The third inequality follows from the usual change of variable and Assumption 2.3.3.e). The second to last line is due to the fact that $\int K(u)du = 1$. The last line follows from the Cauchy-Schwarz inequality and Assumption 2.3.1.b).

To bound Σ_{22}, defined in (A.66) I use Davydov's lemma. Using similar arguments as above it follows that

$$
\begin{aligned}
&\left\| \big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}} - m^{(t)}(X_1)\big) K\Big(\frac{X_1-x}{h_n}\Big) \right\|_\delta^2 \\
&\qquad = \left(\mathbb{E}\left| \Big(\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}} - m^{(t)}(X_1)\Big) K\Big(\frac{X_1-x}{h_n}\Big) \right|^\delta \right)^{2/\delta} \\
&\qquad = \left(\mathbb{E}\left|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}} + C_{11}\right|^\delta \mathbb{E}\left(K^\delta\Big(\frac{X_1-x}{h_n}\Big) \Big| X_2 \right) \right)^{2/\delta} \\
&\qquad \leq \overline{C} h_n^{2\delta} \left(\mathbb{E}\left|\phi(X_2)\mathbb{1}_{\{|\phi(X_2)|\leq\tau_n\}} + C_{11}\right|^\delta \right)^{2/\delta} \\
&\qquad \leq \overline{C} h_n^{2\delta} \left(\mathbb{E}\big(\big|\phi(X_2)\big| + C_{11}\big)^\delta \right)^{2/\delta} \\
&\qquad \leq \overline{C} h_n^{2\delta},
\end{aligned}
$$

for $s \geq \delta > 2$ and suitable constant $\overline{C}$. Note that $s \geq \delta$ is important because otherwise Assumption 2.3.1.b) may not be satisfied. According to Davydov's lemma,

$$
\Sigma_{22} \leq \text{constant}_3 \times r_n h_n \sum_{k=\lfloor h_n^{-1} \rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta}, \qquad \text{(A.68)}
$$

$a > 1-2/\delta$, similar to (A.13). Note that the infinite sum of (A.68) is summable according to Assumption 2.3.4 because $s \geq \delta > 2$. Combining (A.64), (A.65), (A.66), (A.67), and (A.68)

result in

$$\sigma_{r_n}^2 \leq r_n h_n \left(\text{constant}_1 + \text{constant}_2 + \text{constant}_3 \times \sum_{k=\lfloor h_n^{-1} \rfloor}^{\infty} k^a (\alpha(k))^{1-2/\delta} \right).$$

Because the bound b depends on τ_n define $r_n = \lfloor \tau_n^{-1} \sqrt{nh_n / \ln(n)} \rfloor$ and let $\varepsilon = \varepsilon_n = C_\varepsilon \sqrt{\ln(n)/(nh)}$ as before. Then, applying the exponential type inequality in (B.81) to (A.63) results in

$$\begin{aligned} \Lambda_n \sup_{x \in \mathbb{S}} \left\{ \mathbb{P}\left(\left| \sum_{i=1}^{n} U_i(x) \right| > \varepsilon_n n h_n \right) \right\} &\leq 4\Lambda_n \exp\left\{ -\frac{(\varepsilon_n n h_n)^2}{64 \frac{n}{r_n} \sigma_{r_n}^2 + \frac{8}{3} \varepsilon_n n h_n r_n b} \right\} + \frac{4\Lambda_n n}{r_n} \alpha(r_n) \\ &\leq 4\Lambda_n \exp\{-C_\varepsilon \ln(n)\} \\ &\quad + 4n \left(\frac{\tau_n}{h_n} \right)^2 \alpha\left(\left\lfloor \frac{1}{\tau_n} \sqrt{\frac{nh_n}{\ln(n)}} \right\rfloor \right). \end{aligned} \tag{A.69}$$

The first and second term of (A.69) are derived similar to (A.16) and (A.17), respectively. Choosing C_ε sufficiently large and employing condition i) of the lemma guarantees summability of (A.69), i.e., $\sum_{i=1}^{\infty} \mathbb{P}(B_n > C_\varepsilon \sqrt{\ln(n)/(nh_n)}) < \infty$. By virtue of the Borel-Cantelli lemma

$$B_n = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \qquad \text{almost surely.} \tag{A.70}$$

Combing (A.61), (A.62), and (A.70) proves

$$\sup_{x \in \mathbb{S}} \left| T_n^{(t)}(x) \right| = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \quad \text{almost surely,} \tag{A.71}$$

which completes the first part of the lemma.

For the second part of the Lemma note that, given the second term of (A.69) and the definition of τ_n,

$$\begin{aligned} &n \left(\frac{\tau_n}{h_n} \right)^2 \alpha\left(\left\lfloor \frac{1}{\tau_n} \sqrt{\frac{nh_n}{\ln(n)}} \right\rfloor \right) \\ &\qquad = n^{1+2\theta-\beta(1-\theta)/2+(2+\beta)/s} (\ln(n))^{(2+\beta)/s-2\theta+\beta(1-\theta)/2} (\ln(\ln(n)))^{2(2+\beta)/s}. \end{aligned} \tag{A.72}$$

Solving $c = 1 + 2\theta - \beta(1-\theta)/2 + (2+\beta)/s < -1$ for θ, assuming $s > 2$, $\beta > 0$, and $\theta > 0$ leads to $\theta \in (0, (\beta(1-2/s) - 4/s - 4)/(\beta+4))$ with $\beta > 4(1+s)/(s-2)$.

To prove summability note that given $c < -1$, $(2+\beta)/s < \beta(1-\theta)/2$. Thus, $\ln(\ln(n)))^{2(2+\beta)/s} < (\ln(n))^{\beta(1-\theta)}$. Furthermore, the exponent of $\ln(n)$, given in (A.72), simplifies to $(2+\beta)/s - 2\theta + \beta(1-\theta)/2 < \beta(1-\theta)$. Hence, it suffices to prove summability of

$$n^{1+2\theta-\beta(1-\theta)/2+(2+\beta)/s}(\ln(n))^{2\beta(1-\theta)},$$

to establish the result. The integral test shows that

$$\int_e^\infty y^{1+2\theta-\beta(1-\theta)/2+(2+\beta)/s}(\ln(y))^{2\beta(1-\theta)}dy = \int_1^\infty \frac{e^{-a_1 t}}{t^{a_2}}dt < \infty,$$

with $a_1 = -(2+\beta)/s + \beta(1-\theta)/2 - 2 - 2\theta$ and $a_2 = -4\beta(1-\theta)$. Note that $a_2 < -1$ and $a_1 > 0$ because, given the definition of c, $(2+\beta)/s - \beta(1-\theta)/2 < -2-2\theta$ and therefore $-(2+\beta)/s + \beta(1-\theta)/2 > 2+2\theta$. Using $h_n \sim (\ln(n)/n)^\theta$ and (A.71) it is easy to see that

$$\sup_{x\in\mathbb{S}} \left| T_n^{(t)}(x) \right| = \mathcal{O}\left(\left(\frac{\ln(n)}{nh_n} \right)^{(1-\theta)/2} \right) \quad \text{almost surely,}$$

which completes the proof.

□

A.11 Proof of Proposition 2.3.15

Due to Lemma 2.3.13 define the truncated version of $J_1(x)$ as

$$J_1^{(t)}(x) = \sum_{i=1}^n \Big(\phi(X_{i+1}) \mathbb{1}_{\{|\phi(X_{i+1})| \le \tau_n\}} - m^{(t)}(X_i) \Big) p_i(x;\lambda_n) K_{h_n}(X_i - x).$$

Thus, proving

$$\sup_{x\in\mathbb{S}} \left| J_1^{(t)}(x) \right| = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}} \right) \qquad \text{almost surely,}$$

proves the statement of the proposition. Note that

$$\begin{aligned}
\sup_{x\in\mathbb{S}} \left| J_1^{(t)}(x) \right| &= \sup_{x\in\mathbb{S}} \left| \sum_{i=1}^n \Big(\phi(X_{i+1}) \mathbb{1}_{\{|\phi(X_{i+1})| \le \tau_n\}} - m^{(t)}(X_i) \Big) \left(p_i(x;\lambda_n) - \frac{1}{n} \right) K_{h_n}(X_i - x) \right| \\
&\quad + \sup_{x\in\mathbb{S}} \left| \frac{1}{n} \sum_{i=1}^n \Big(\phi(X_{i+1}) \mathbb{1}_{\{|\phi(X_{i+1})| \le \tau_n\}} - m^{(t)}(X_i) \Big) K_{h_n}(X_i - x) \right| \\
&=: A_n + B_n. \qquad \text{(A.73)}
\end{aligned}$$

Similar to the arguments used in the proofs of Proposition 2.3.8 and 2.3.10, A_n is bounded as follows,

$$\begin{aligned} A_n &\leq \sup_{x\in\mathbb{S}}\{\big|C_4\lambda_n(x)\big|\}\sup_{x\in\mathbb{S}}\Bigg\{\Bigg|\frac{1}{n}\sum_{i=1}^{n}\Big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}-m^{(t)}(X_i)\Big) \\ &\qquad\qquad \times\left(\frac{1}{1+\lambda_n(x)(X_i-x)K_{h_n}(X_i-x)}\right)K_{h_n}(X_i-x)\Bigg|\Bigg\} \\ &=: A_{n,1}A_{n,2}. \end{aligned} \tag{A.74}$$

The term $A_{n,1}$ is equivalent to (A.28), thus, according to (A.29),

$$A_{n,1} = \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \text{almost surely.} \tag{A.75}$$

To determine $A_{n,2}$ note that, for n sufficiently large,

$$\begin{aligned} A_{n,2} &= \frac{1}{1-\Delta_n}\sup_{x\in\mathbb{S}}\left|\frac{1}{n}\sum_{i=1}^{n}\Big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}-m^{(t)}(X_i)\Big)K_{h_n}(X_i-x)\right| \\ &=: A_{n,21}A_{n,22}, \end{aligned} \tag{A.76}$$

where (A.30) is used. The first term, $A_{n,21}$, is equivalent to (A.31), thus, according to (A.32),

$$A_{n,21} = 1 + \mathcal{O}(h_n) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \quad \text{almost surely.} \tag{A.77}$$

Regarding the second term of (A.76) note that

$$\begin{aligned} A_{n,22} &= \sup_{x\in\mathbb{S}}\left|\frac{1}{n}\sum_{i=1}^{n}\Big(\phi(X_{i+1})\mathbb{1}_{\{|\phi(X_{i+1})|\leq\tau_n\}}-m^{(t)}(X_i)\Big)K_{h_n}(X_i-x)\right| \\ &= \sup_{x\in\mathbb{S}}\big|\,T_n^{(t)}(x)\big|. \end{aligned}$$

According to Lemma 2.3.14,

$$A_{n,22} = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \tag{A.78}$$

Inserting (A.77) and (A.78) into (A.74) lead to

$$A_n = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \tag{A.79}$$

Since B_n, given in (A.73), is equivalent to $A_{n,22}$ it follows that

$$B_n = \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh_n}}\right) \qquad \text{almost surely.} \tag{A.80}$$

Combing (A.73), (A.79), and (A.80) proves the statement.

□

A.12 Proof of Theorem 2.3.16

I rewrite (2.12) in order to apply the previous results. It follows that,

$$\begin{aligned}\sup_{x\in\mathbb{S}}\left|\widehat{m}(x) - m(x)\right| &= \sup_{x\in\mathbb{S}}\left|\frac{J_1(x)}{J_3(x)} + \frac{J_2(x)}{J_3(x)}\right| \\ &\leq \sup_{x\in\mathbb{S}}\left\{\left|J_1(x)\right|\right\}\sup_{x\in\mathbb{S}}\left\{\left|\frac{1}{J_3(x)}\right|\right\} + \sup_{x\in\mathbb{S}}\left\{\left|J_2(x)\right|\right\}\sup_{x\in\mathbb{S}}\left\{\left|\frac{1}{J_3(x)}\right|\right\}.\end{aligned}$$

Using Corollaries 2.3.9 and 2.3.11 and Proposition 2.3.15 proves the first part of the theorem.

To prove the second part use $h_n \sim (\ln(n)/n)^\theta$, with $\theta \in (0,1)$, and the fact that $\ln(n)/n \to 0$ to determine the order.

□

B Additional lemmas

Lemma B.1. *Let $\{Z_i\}_{i=-\infty}^{\infty}$ be a stationary zero-mean real-valued α-mixing process such that $\mathbb{P}(|Z_i| \leq b) = 1$. Then for each integer $s_n \leq n$ and η such that $s_n < \eta b/4$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i\right| > \eta\right) \leq 4\exp\left\{-\frac{\eta^2}{64\frac{n}{s_n}\sigma_{s_n}^2 + \frac{8}{3}\eta s_n b}\right\} + 4\frac{n}{s_n}\alpha(s_n) \tag{B.81}$$

where

$$\sigma_{s_n}^2 = \mathbb{E}\left(\sum_{i=1}^{s_n} Z_i\right)^2. \tag{B.82}$$

Proof. The lemma is due to Liebscher (1996, Theorem 2.1) and is a direct consequence of Rio (1995, Theorem 5). A proof can be found in Rio (1995).

□

Lemma B.2 (Davydov's lemma). *Let X and Y be two random variables defined a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let α be the strong mixing coefficient measuring the dependence of X and Y. If $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|Y|^q < \infty$ for $p, q \geq 1$ and $1/p + 1/q < 1$, it follows that*

$$\left|cov(X, Y)\right| \leq 8\alpha^{1-1/p-1/q}\|X\|_p\|Y\|_q,$$

with $\|X\| = \left(\mathbb{E}\left|X\right|^p\right)^{1/p}$ and $\|Y\| = \left(\mathbb{E}\left|Y\right|^q\right)^{1/q}$.

Remark B.3. The original statement of the corollary following Lemma 2.1 in Davydov (1968) the constant 12 instead of 8 is used.

Proof. See Hall and Heyde (1980, p. 278).

□

References

Bao, Y., T.-H. Lee, and B. Saltoğlu (2006): "Evaluating Predictive Performance of Value-at-Risk Models in Emerging Markets: A Reality Check," *Journal of Forecasting*, 25, 101–128.

Basrak, B., R. A. Davis, and T. Mikosch (2002): "Regular Variation of GARCH Processes," *Stochastic Processes and their Applications*, 99, 95–115.

Bradley, R. C. (2005): "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions," *Probability Surveys*, 2, 107–144.

Cai, Z. (2002): "Regression Quantiles for Time Series," *Econometric Theory*, 18, 169–192.

Cai, Z. and X. Wang (2008): "Nonparametric Estimation of Conditional VaR and Expected Shortfall," *Journal of Econometrics*, 147, 120–130.

Chen, S. X. and P. Hall (1993): "Smoothed Empirical Likelihood Condfidence Intervals for Quantiles," *The Annals of Statistics*, 21, 1166–1181.

Davidson, J. (1994): *Stochastic Limit Theory*, Oxford: Oxford University Press.

Davydov, Y. A. (1968): "Convergence of Distributions Generated by Stationary Stochastic Processes," *Theory of Probability & Its Applications*, 13, 691–696.

Efron, B. and R. J. Tibshirani (1994): *An Introduction to the Bootstrap*, Boca Raton: Chapman & Hall/CRC.

Fan, J., T. Gasser, I. Gijbels, M. Brockmann, and J. Engel (1995): "On Nonparametric Estimation via Local Polynomial Regression," *Working paper University of Louvain.*

Fan, J. and I. Gijbels (1996): *Local Polynomial Modelling and its Applications*, Boca Raton: Chapman & Hall/CRC.

Fan, J. and Q. Yao (2005): *Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer.

Fryzlewicz, P. and S. Subba Rao (2011): "Mixing Properties of ARCH and Time-Varying ARCH Processes," *Bernoulli*, 17, 320–346.

Hall, P. and C. C. Heyde (1980): *Martingale Limit Theory and its Application*, New York: Academic Press.

Hansen, B. E. (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, 24, 726–748.

Kato, K. (2012): "Weighted Nadaraya-Watson Estimation of Conditional Expected Shortfall," *Journal of Financial Econometrics*, 10, 265–291.

Kristensen, D. (2009): "Uniform Convergence Rates of Kernel Estimators with Heterogeneous Dependent Data," *Econometric Theory*, 25, 1433–1445.

Lahiri, S. N. (2003): *Resampling Methods for Dependent Data*, New York: Springer.

Li, Q. and S. Racine (2006): *Nonparametric Econometrics. Theory and Practice*, Princeton: Princeton University Press.

Liebscher, E. (1996): "Strong Convergence of Sums of α-Mixing Random Variables with Applications to Density Estimation," *Stochastic Processes and their Applications*, 65, 69–80.

Mack, Y. P. and B. W. Silverman (1982): "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61, 405–415.

Masry, E. (1996): "Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates," *Journal of Time Series Analysis*, 17, 571–599.

Paparoditis, E. and D. N. Politis (2000): "The Local Bootstrap for Kernel Estimators under General Dependence Conditions," *Annals of the Institute of Statistical Mathematics*, 52, 139–159.

Rio, E. (1995): "The Functional Law of the Iterated Logarithm for Stationary Strongly Mixing Sequences," *Annals of Probability*, 23, 1188–1203.

Rosenblatt, M. (1956): "A Central Limit Theorem and a Strong Mixing Condition," *Proceedings of the National Academy of Sciences of the United States of America*, 42, 43–47.

Steikert, K. U. (2014a): "A Local Bootstrap Procedure to Select the Bandwidth for the Weighted Nadaraya-Watson Estimator in Case of Weakly Dependent Data," *Working paper University of Zurich.*

——— (2014b): "The Weighted Nadaraya-Watson Estimator: Pointwise Strong Consistency and Convergence Rates for Strongly Mixing Processes," *Working paper University of Zurich.*

Stone, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053.

Tay, A. S. and C. Ting (2008): "Intraday Stock Prices, Volume, and Duration: A Nonparametric Conditional Density Analysis," in *High frequency financial econometrics: recent developments*, ed. by L. Bauwens, W. Pohlmeier, and D. Veredas, Heidelberg: Physica-Verlag, 253–268.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge: Cambridge University Press.

Chapter 3

A Local Bootstrap Procedure to Select the Bandwidth for the Weighted Nadaraya-Watson Estimator in Case of Weakly Dependent Data

3.1 Introduction

An eminent aspect in the study of nonlinear times series is nonlinear forecasting, in particular forecasting future events, such as financial returns of some asset, given a selective past in the form of lagged variables. Estimating the expected future event given a selective past (henceforth prediction), is only the beginning in a serious attempt to forecast because the predictive distribution function, i.e., the distribution of the future event conditional on the selective past, contains all necessary information about this future event. Correctly estimating this conditional cumulative distribution function (CDF) is therefore essential to forecast time series.

Previous standard nonparametric estimation attempts fail in providing a coherent forecasting framework with favorable bias properties. Although the Nadaraya-Watson estimator also provides a coherent forecasting framework, i.e., prediction and predictive distribution can properly be estimated, the estimator exhibits inferior bias properties to the local linear estimator. The local linear estimator, however, produces only proper prediction estimates while for estimating the predictive distribution negative probabilities cannot be ruled out. In response to the desire to construct an estimator with the same bias properties as the local linear estimator while preserving the property that the Nadaraya-Watson estimator is always a proper estimator of the conditional CDF, Hall et al. (1999) propose the weighted Nadaraya-Watson estimator. Their proposal and the generalizations of Cai (2001) provide a coherent forecasting framework. That is, prediction and the predictive distribution can properly be estimated with both exhibiting the same favorable bias property. For these estimates selecting the bandwidth is important and it is therefore absolutely necessary to have sound procedures at hand to select this free parameter for this particular estimator.

The selection of the bandwidth for nonparametric estimators such as the weighted Nadaraya-Watson estimator and the class of local polynomial estimators is a crucial task of the estimation process. This is because the bandwidth, being a free parameter, strongly influences the resulting estimates. A small value of the bandwidth implies a more volatile estimate in the sense that structures belonging exclusively to the observed data, and not to the underlying signal, may have a larger impact on the estimate. A large value, on the other hand, increases the bias of the estimate by over-smoothing the data such that essential parts of the signal is smoothed away. The fundamental task of the bandwidth is therefore to balance the bias and variance of the estimator. Although the estimators also depend on the

choice of the kernel function, bandwidth selection remains more important, because Fan et al. (1995, Theorem 1) show that the Epanechnikov kernel is optimal in a minimax sense for local polynomial regression estimators.

There are two prominent alternatives for bandwidth selection for the weighted Nadaraya-Watson estimator in case of time series data. The first is the bootstrap procedure proposed in Hall et al. (1999). Their procedure is based on the estimation of a parametric data generating process. The bootstrap sample is then recursively constructed using the fitted model and the (approximately) independent and identically distributed (iid) centered residuals. For details on the, so called, autoregressive bootstrap (ARB) see for example Lahiri (2003, pp. 199–220). Employing the ARB to select the bandwidth however heavily depends on the estimation of the data generating process. Regarding the ARB, Politis, Romano, and Wolf (1999, p. 65) note: "[. . .] this approach is restricted to situations where a general regression model can be relied upon" and Lahiri (2003, p. 25) adds: "[i]n a problem where the statistician does not have enough prior knowledge to specify such models, these models are not very useful." A close alternative is presented by Fan and Yao (2005, pp. 457–458). They suggest to fit a simple linear autoregressive process to the data and use the fitted model for the ARB. The linearity of the autoregressive model constitutes a strong limitation and will lead to unsatisfactory results for nonlinear data. The second alternative is to select the bandwidth according to a nonparametric version of Akaike's information criterion (AIC). The method is proposed in Cai and Tiwari (2000) and finds application in Cai (2002). Their method, which is introduced in Section 3.3, will serve as a benchmark method in the simulations study below.

This manuscript proposes a new, fully data driven method to select the bandwidth for the weighted Nadaraya-Watson estimator. The method requires only weak assumptions, namely, that data are assumed to be strictly stationary and weakly dependent in the form of strongly mixing time series. Examples include (types of) ARMA, Markov, and GARCH processes. The procedure is based on the local bootstrap proposed by Paparoditis and Politis (2000). They develop the procedure to approximate the sampling distribution of kernel estimators, in particular they consider the Nadaraya-Watson estimator. For the present manuscript their approach is extended to the weighted Nadaraya-Watson estimator due to its favorable bias properties. The procedure extends because the estimator consistently estimates the predictive distribution.[1] Moreover, because the limiting distribution of the weighted

[1]For consistency results see Cai (2001) and Steikert (2014).

Nadaraya-Watson estimator is Gaussian, depending on the marginal distribution of the data and the predictive distribution (see Cai (2001, Theorem 1)). The bootstrap procedure is able to consistently estimate theses unknowns.

Given the local bootstrap an estimator of the integrated mean squared error (IMSE) is constructed and the bandwidth is selected such that it minimizes the estimated IMSE. The IMSE is a convenient function to determine the bandwidth because the mean squared error is equivalent to the sum of squared bias and the variance of the weighted Nadaraya-Watson estimator. The selection of the bandwidth therefore balances the bias and variance of the estimator on the entire set of evaluation points (locations). Due to the consideration of the local bootstrap an additional bandwidth, the so called pilot or resampling bandwidth, emerges. The selection of bandwidth therefore depends on the choice of the pilot bandwidth. The dependence issue is dampened by introducing an iterated bandwidth selection scheme and provide simple choices for the initial pilot bandwidth.

The advantage of the method are twofold. Given the selective past only future observations are bootstrapped. It is therefore easy to implement and capable of coping with large samples as well as with a large number of bootstrap samples to approximate the relevant bootstrap statistic. Its eminent feature is that it is entirely nonparametric; it therefore avoids any assumption regarding the parametric form of the data generating process. This is appealing because in most cases, at least if nonparametric estimators are considered, the observed data may be generated by a nonlinear process implying cumbersome techniques to estimate this process.

To measure the performance of the proposed method to select the bandwidth for the weighted Nadaraya-Watson estimator, an extensive simulation study is employed. For various data generating processes, such as the exponential autoregressive (AR), smooth transition AR, polynomial AR, rational nonlinear AR, and the autoregressive conditional heteroscedastic models, I consider one-and three step ahead predictions of the time series as well as the variance, using the weighted Nadaraya-Watson estimator and selecting the bandwidth according to the proposed procedure. To compare the results and taking advantage of the simulation study framework the empirically optimal bandwidth is computed. This bandwidth is optimal in a squared sense, minimizing the squared difference of the prediction estimates and the value being estimated on a given set of locations. Performance of the selection procedure is measured by the mean absolute deviation error (MADE) which is commonly used in the literature. It measures, for a given bandwidth, the mean absolute difference of the prediction estimates and the value being estimated on a given set of locations.

The results of the simulation study indicate a selection of the bandwidth, such that the MADEs, given the bandwidth selected via the local bootstrap procedure, are close to the MADEs given the empirically optimal bandwidth. This implies that the proposed local bootstrap procedure is an appealing choice among the scarce list of bandwidth selection methods for the weighted Nadaraya-Watson estimator. The benchmark method, based on a nonparametric version of AIC, is outperformed in every example considered in this study.

The remainder of the manuscript is organized as follows. Section 3.2 introduces the Nadaraya-Watson, local linear, and weighted Nadaraya-Watson estimators. The section also compares the estimators in terms of bias and effective kernel weights and highlights the features of each estimator. The discussion reveals that the weighted Nadaraya-Watson estimator combines favorable properties of the Nadaraya-Watson and local linear estimators. Section 3.3 briefly repeats the bandwidth selection method based on a nonparametric version of Akaike's information criterion. In Section 3.4 the local bootstrap procedure is introduced and an iterative algorithm is presented to select the bandwidth for the weighted Nadaraya-Watson estimator. The selection of bandwidth is tested in an extensive simulation study with the most common nonlinear times series models. The results of this and the benchmark method are given in Section 3.5. A summary is given in Section 3.6 while the appendix contains all tables and figures.

3.2 The estimators

This section introduces the Nadaraya-Watson, local linear, and weighted Nadaraya-Watson estimators. Advantages as well as deficits of the estimators are discussed in detail. In particular the bias properties are compared. The discussion will show that the weighted Nadaraya-Watson estimator is the only estimator of the three providing a coherent approach to estimate all statistics necessary for forecasting nonlinear time series.

Let $\{X_t\}_{t=1}^T$ denote a strictly stationary real-valued time series. Furthermore, let $\phi(\cdot)$ denote an arbitrary Borel-measurable function on $\mathbb{R}$ and assume $\mathbb{E}\,|\phi(X_{t+1})| < \infty$. Consider the following nonparametric regression model

$$\phi(X_{t+1}) = m(\boldsymbol{X}_t) + \sigma(\boldsymbol{X}_t)\epsilon_{t+1}, \tag{3.1}$$

with $\boldsymbol{X}_t = (X_{t-i_1}, X_{t-i_2}, ..., X_{t-i_d})$ denoting the d-dimensional vector of lagged variables, with integers $i_1, ..., i_d$ satisfying $0 \leq i_1 < i_2 < ... < i_d < \infty$. The errors ϵ_{t+1} are independent of $\{X_k\}_{k \leq t}$, satisfying $\mathbb{E}(\epsilon_{t+1}|\boldsymbol{X}_t) = 0$ and $\text{var}(\epsilon_{t+1}|\boldsymbol{X}_t) = 1$. The functional form of $m(\cdot)$ is unknown but by assuming sufficient smoothness it can be estimated nonparametrically. For this define the regression function $m(\boldsymbol{x})$ as the conditional mean of $\phi(X_{t+1})$ given the selective past $\boldsymbol{X}_t = \boldsymbol{x}$, i.e.,

$$m(\boldsymbol{x}) = \mathbb{E}\big(\phi(X_{t+1})\big|\boldsymbol{X}_t = \boldsymbol{x}\big), \tag{3.2}$$

with $\boldsymbol{x} = (x_{i_1}, ..., x_{i_d})$. The regression function in (3.2) represents the best mean squared prediction of $\phi(X_{t+1})$ based on the information $\boldsymbol{X}_t = \boldsymbol{x}$ (see Li and Racine (2006, p. 59)). Introducing $\phi(\cdot)$ as a response function provides an enriched flexibility in expressing various time series statistics of the data such as l-step ahead predictions ($\phi(X_{i+1}) = X_{i+l}$, with $i_1 = l - 1$ and $d = 1$), raw moments thereof ($\phi(X_{t+l}) = X_{t+l}^k$ for integers $k > 0$), or conditional probabilities ($\phi(X_{t+l}) = \mathbb{1}_{(-\infty,y]}(X_{t+l})$ for some $y \in \mathbb{R}$).

To estimate the function $m(\cdot)$ nonparametrically using the Nadaraya-Watson, local linear, and weighted Nadaraya-Watson estimators the location $\boldsymbol{x}$ provides very little information about $m(\boldsymbol{x})$. By including data of lagged variables, i.e., $\{\boldsymbol{X}_t\}_{t=i_d+1}^{T-1}$, in a close neighborhood of $\boldsymbol{x}$ more information about $m(\cdot)$ is revealed. Presuming the existence of the $(q+1)$-th derivate of $m(\cdot)$ locally at $\boldsymbol{x}$ and approximating $m(\boldsymbol{X}_t)$ in the local neighborhood of $\boldsymbol{x}$ by a multivariate polynomial of order q leads to

$$\begin{aligned} m(\boldsymbol{X}_t) &\approx \sum_{j=0}^{q} \sum_{\substack{\alpha_1=0 \\ \alpha_1+\cdots+\alpha_d=j}}^{j} \cdots \sum_{\alpha_d=0}^{j} \frac{D^{(\alpha_1,...,\alpha_d)} m(\boldsymbol{x})}{\alpha_1! \cdots \alpha_d!} (X_{t-i_1} - x_{i_1})^{\alpha_1} \cdots (X_{t-i_d} - x_{i_d})^{\alpha_d} \\ &= \sum_{0 \leq |\boldsymbol{\alpha}| \leq q} \frac{D^{\boldsymbol{\alpha}} m(\boldsymbol{x})}{\boldsymbol{\alpha}!} (\boldsymbol{X}_t - \boldsymbol{x})^{\boldsymbol{\alpha}}, \end{aligned} \tag{3.3}$$

where $D^{\boldsymbol{\alpha}} m(\boldsymbol{x})$ abbreviates $D^{\boldsymbol{\alpha}} m(\boldsymbol{v}) = (\partial^{|\boldsymbol{\alpha}|} m(\boldsymbol{v}) / (\partial v_1^{\alpha_1} \cdots \partial v_d^{\alpha_d}))|_{\boldsymbol{v}=\boldsymbol{x}}$ with the (usual) multi-index $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_d)$ and notation $|\boldsymbol{\alpha}| = \alpha_1 + ... + \alpha_d$, $\boldsymbol{\alpha}! = \alpha_1! \cdots \alpha_d!$, and $\boldsymbol{v}^{\boldsymbol{\alpha}} = v_1^{\alpha_1} \cdots v_d^{\alpha_d}$, for any vector $\boldsymbol{v}$. This polynomial is fitted locally by a weighted

multivariate polynomial regression, i.e.,

$$\sum_{t=i_d+1}^{T-1}\left(\phi(X_{t+1})-\sum_{0\leq|\boldsymbol{\alpha}|\leq q}\beta_{\boldsymbol{\alpha}}(\boldsymbol{x})(\boldsymbol{X}_t-\boldsymbol{x})^{\boldsymbol{\alpha}}\right)^2 p_t(\boldsymbol{x};\boldsymbol{\lambda})K(h^{-1}(\boldsymbol{X}_t-\boldsymbol{x})), \tag{3.4}$$

is minimized with respect to $\beta_{\boldsymbol{\alpha}}$ leading to estimates $\widehat{\beta}_{\boldsymbol{\alpha}}$. By (3.3), $\widehat{\beta}_{\boldsymbol{\alpha}}$ estimates $D^{\boldsymbol{\alpha}}m(\boldsymbol{x})/(\boldsymbol{\alpha}!)$ and therefore $\widehat{D^{\boldsymbol{\alpha}}m}(\boldsymbol{x}) = \boldsymbol{\alpha}!\widehat{\beta}_{\boldsymbol{\alpha}}$. In particular I focus on $\widehat{D^{(0,\ldots,0)}m}(\boldsymbol{x}) = \widehat{m}(\boldsymbol{x})$.[2] The functions $K : \mathbb{R}^d \to \mathbb{R}_+$ and $p_t : \mathbb{R}^d \to \mathbb{R}_{++}$ are both weight functions. The kernel function $K(\cdot)$ is assumed to be a spherically symmetric density satisfying $\int u_l K(\boldsymbol{u})d\boldsymbol{u} = 0$, $\int u_l u_k K(\boldsymbol{u})d\boldsymbol{u} = 0$, for $k, l = 1, ..., d$ and $k \neq l$, as, e.g., the spherically Epanechnikov kernel function (see Fan and Yao (2005, p. 315)). The integrals indicate multivariate integration over the d-dimensional Euclidean space, i.e., $\int K(\boldsymbol{u})d\boldsymbol{u} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} K(u_1, ..., u_d)du_1\cdots du_d$. $K(\cdot)$ controls the amount of information of each $\boldsymbol{X}_t$ relevant to estimate $m(\boldsymbol{x})$ with the bandwidth h controlling the size of the local neighborhood around $\boldsymbol{x}$. For unimodal symmetric kernels the contributions of the data, to determine $\widehat{m}(\boldsymbol{x})$, for which the Euclidean distance between $\boldsymbol{X}_t$ and $\boldsymbol{x}$ is large is down-weighted, opposed to data for which $\boldsymbol{X}_t$ and $\boldsymbol{x}$ are close. The probabilities $p_t(\boldsymbol{x};\boldsymbol{\lambda})$ are either uniformly constant or location dependent contingent on the particular estimator under study.

Because of the approximation in (3.3), the resulting nonparametric estimators exhibit finite-sample bias, i.e., $\mathbb{E}\,\widehat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) \neq 0$. It is easy to see that there exists a relation between the bias and the precision of (3.3). Increasing the order of the polynomial of the approximation and/or decreasing the bandwidth h decreases the bias because the approximation is more accurate. This decrease via q and h, however, increases the variability of the estimator. Thus, there is a tradeoff between bias and variance that needs to be balanced to obtain a smooth estimated surface with low bias and low variance. In addition, for certain small values of the bandwidth the approximation may not exist. For an illustration suppose $d = q = 1$, then at least two observations must lie in the neighborhood of x to guarantee the existence of $\widehat{m}(x)$.

3.2.1 The Nadaraya-Watson estimator

To introduce the *ordinary* Nadaraya-Watson estimator suppose that the probabilities $p_t(\boldsymbol{x};\boldsymbol{\lambda})$ are uniform, i.e., $p_t(\boldsymbol{x};\boldsymbol{\lambda}) = (T-i_d-1)^{-1}$ for all $t = i_d+1, i_d+2, ..., T-1$.

[2] For derivative estimators see Masry (1996a) and Masry (1996b).

Then (3.4) reduces to the well known optimization problem to determine the local polynomial estimators. If the approximation in (3.3) is constant, i.e., $q = 0$ implying $\boldsymbol{\alpha} = (0, ..., 0)$, then $m(\boldsymbol{x})$ is estimated by the ordinary Nadaraya-Watson estimator which reads

$$\widehat{m}_{nw}(\boldsymbol{x}) = \frac{\sum_{t=i_d+1}^{T-1} K(h^{-1}(\boldsymbol{X}_t - \boldsymbol{x}))\phi(X_{t+1})}{\sum_{t=i_d+1}^{T-1} K(h^{-1}(\boldsymbol{X}_t - \boldsymbol{x}))}. \tag{3.5}$$

This estimator, introduced independently by Nadaraya (1964) and Watson (1964), features simplicity and monotonicity in y if $\phi(X_{t+1}) = \mathbb{1}_{(-\infty,y]}(X_{t+1})$, for some $y \in \mathbb{R}$. This is because the kernel function $K(\cdot)$ is usually assumed to be non-negative. Thus, the ordinary Nadaraya-Watson estimator is a proper estimator of the conditional CDF. A major drawback, however, is the inferior bias compared to higher-order polynomial estimators. To compare the bias of a given estimator I focus on the non-stochastic part of the bias and name this part simply the bias (see Gu et al. (2013, p. 3) for a formal definition). The bias of the Nadaraya-Watson estimator is given by

$$\begin{aligned}\text{bias}(\widehat{m}_{nw}(\boldsymbol{x})) = \frac{h^2}{2} \sum_{|\boldsymbol{\alpha}|=1} \int (\boldsymbol{u}^{\boldsymbol{\alpha}})^2 K(\boldsymbol{u}) d\boldsymbol{u} \Bigg(& D^{2\boldsymbol{\alpha}} m(\boldsymbol{x}) \\ & + \frac{2}{f_{\boldsymbol{X}_t}(\boldsymbol{x})} (D^{\boldsymbol{\alpha}} m(\boldsymbol{x}))(D^{\boldsymbol{\alpha}} f_{\boldsymbol{X}_t}(\boldsymbol{x})) \Bigg),\end{aligned} \tag{3.6}$$

where $\sum_{|\boldsymbol{\alpha}|=1} D^{2\boldsymbol{\alpha}}(\boldsymbol{x})$ is the trace of the $(d \times d)$-dimensional Hessian matrix of $m(\boldsymbol{x})$ and $f_{\boldsymbol{X}_t}(\cdot)$ denotes the (marginal) density of $\boldsymbol{X}_t$. I omit a variance comparison of the estimators considered because all three exhibit the same variance given by

$$\frac{\sigma^2(\boldsymbol{x}) \int K^2(\boldsymbol{u}) d\boldsymbol{u}}{T h^d f_{\boldsymbol{X}_t}(\boldsymbol{x})}. \tag{3.7}$$

3.2.2 The local linear estimator

A linear approximation in (3.3), i.e., $q = 1$ as well as uniform probabilities lead to the local linear estimator which reads

$$\widehat{m}_{ll}(x) = \frac{\sum_{t=1}^{T} w_t(x)\phi(X_{t+1})}{\sum_{t=1}^{T} w_t(x)}, \tag{3.8}$$

with weights

$$w_t(x) = K\left(\frac{X_t - x}{h}\right)\left(\sum_{k=1}^{T}(X_k - x)^2 K\left(\frac{X_k - x}{h}\right)\right.$$
$$\left. -(X_t - x)\sum_{k=1}^{T}(X_k - x)K\left(\frac{X_k - x}{h}\right)\right), \qquad (3.9)$$

for $i_d = 0$. I omit a multivariate representation of the estimator because it requires a substantial amount of additional notation not used further in this manuscript. For details see Masry (1996a, pp. 573–575).

The local linear estimator, introduced by Stone (1977) and Cleveland (1979) and subsequently investigated by Fan and Gijbels (1992), Fan (1993), Ruppert and Wand (1994), and Masry (1996b), among others, exhibits numerous advantages to the Nadaraya-Watson estimator of which only a few are covered here (see Hastie and Loader (1993) and Fan and Gijbels (1996, pp. 60–76) for more details). A major advantage is the favorable bias property due to the linear approximation in (3.3). The bias of $\widehat{m}_{ll}(\boldsymbol{x})$ according to Masry (1996b, p. 95) reads

$$\text{bias}(\widehat{m}_{ll}(\boldsymbol{x})) = \frac{h^2}{2}\sum_{|\boldsymbol{\alpha}|=1} D^{2\boldsymbol{\alpha}} m(\boldsymbol{x}) \int (\boldsymbol{u}^{\boldsymbol{\alpha}})^2 K(\boldsymbol{u}) d\boldsymbol{u}. \qquad (3.10)$$

The difference compared to the bias of the Nadaraya-Watson estimator is the additional term in (3.6), which depends on the derivatives of $m(\boldsymbol{x})$ and $f_{\boldsymbol{X}_t}(\boldsymbol{x})$, not present in (3.10). This term has serious consequences for the bias. For an illustration consider estimating $\mathbb{E}(X_t - x|X_t = x)$ using (3.5). Then,

$$\widehat{m}_{nw}(x) = \frac{\sum_{t=1}^{T} K((X_t - x)/h)(X_t - x)}{\sum_{t=1}^{T} K((X_t - x)/h)} \neq 0, \qquad (3.11)$$

unless the design data is equally spaced around x, due to the symmetry of the kernel function. Thus, whenever the data is not equally spaced around x the estimator is biased. This bias, however, is gratuitous because using the local linear estimator in (3.8) for this example results in

$$\widehat{m}_{ll}(x) = \frac{\sum_{t=1}^{T} w_t(x)(X_t - x)}{\sum_{t=1}^{T} w_t(x)}. \qquad (3.12)$$

Given the weights in (3.9) it is easy to see that $\widehat{m}_{ll}(x) = 0$. In particular, the

bias of the Nadaraya-Watson estimator is particular severe at boundaries of the support of the marginal density $f_{X_t}(x)$. If, e.g., x is equal to the lower bound of the support of $f_{X_t}(x)$, then each $(X_t - x)$ in (3.11) is nonnegative implying an even larger numerator and therefore a larger bias. A further issue is illustrated by assuming a purely linear data generating process, i.e., a deterministic (without noise) relation such that $m(X_t) = a_1 + a_2 X_t$. In this case the Nadaraya-Watson estimator yields a nonlinear function due to the additional bias component. For the local linear estimator the bias is zero because $m''(x) = 0$ but since $m'(x) = a_1 \neq 0$ the Nadaraya-Watson estimator is biased (for a detailed discussion see Chu and Marron (1991, pp. 414–417)).

Fan (1993, pp. 198–199) shows that

$$\frac{1}{T}\sum_{t=1}^{T}(X_t - x)w_t(x) = 0, \tag{3.13}$$

is the reason for the smaller bias of $\widehat{m}_{ll}(x)$. An equivalent condition for the ordinary Nadaraya-Watson estimator would read

$$\frac{1}{T}\sum_{t=1}^{T}(X_t - x)K\left(\frac{X_t - x}{h}\right) = 0, \tag{3.14}$$

which is similar to the numerator of (3.11). As noted above this property is fulfilled only if the observations are equally spaced around x. This case, however, is unlikely to occur for random designs. In summary, the local linear estimator is superior to the Nadaraya-Watson estimator with respect to bias and the reason for this is the discrete moment condition in (3.13).

A major drawback of the local linear estimator are the non-distributional properties once the response function reads $\phi(X_{t+1}) = \mathbb{1}_{(-\infty,y]}(X_{t+1})$ for some $y \in \mathbb{R}$ because the probabilities are not necessarily non-negative since the estimator is not monotone in y. Yu and Jones (1998) as well as in Hall et al. (1999) discus this case in more detail. For an illustration suppose $i_d = 0$, $h = 1$, and $x = 0$, with x being equal to the lower bound of the support of the marginal density $f_{X_t}(x)$ implying $X_t \geq 0$ for all $t = 1, ..., T$. In addition, let $K(\cdot)$ be a standardized kernel function with bounded support such that $K(X_t - 0) = 0$ for all $|X_t| > 1$. An example of such kernel function is the standardized Epanechnikov kernel function (see Subsection 3.5.1). The weights of the local linear estimator at $x = 0$ for this illus-

tration simplify to

$$w_t(0) = K(X_t)\left(\sum_{k=1}^{T-1} X_k^2 K(X_k) - X_t \sum_{k=1}^{T-1} X_k K(X_k)\right).$$

Suppose X_s is the largest observation at the right end of the support of the kernel function at $x = 0$. For example, $X_s = 1 - \varepsilon$, for some $\varepsilon > 0$. Then, $w_s(0) = -K(X_s)\sum_{k=1}^{T-1} X_k(X_s - X_k)K(X_k) < 0$ because $X_k \leq X_t$ and $0 \leq X_k < 1$ for all $k = 1, ..., T-1$. Negative weights cause a non-monotone behavior of the conditional CDF and therefore result in negative probabilities. Because cases like this can not be excluded the local linear estimator is not a proper estimator of the conditional CDF.

In summary, the local linear estimator should be preferred to the Nadaraya-Watson estimator given the favorable bias properties, but once estimating the conditional CDF the local linear estimator should be avoided. It is therefore desirable to design an estimator similar to the Nadaraya-Watson estimator, but at the same time fulfilling a condition equivalent to (3.14), such that the newly designed estimator is a proper estimator of the conditional CDF with the same bias properties as the local linear estimator.

3.2.3 The weighted Nadaraya-Watson estimator

In response to the drawback of the classical methods Hall and Presnell (1999) introduce the *weighted* Nadaraya-Watson estimator combining the favorable features of both aforementioned estimators. This constrained estimator uses the probabilities, $p_t(\boldsymbol{x}; \boldsymbol{\lambda})$ in (3.4), to guarantee an equivalent condition as the one given in (3.14), i.e., it enforces

$$\sum_{t=i_d+1}^{T-1} (\boldsymbol{X}_t - \boldsymbol{x}) p_t(\boldsymbol{x}; \boldsymbol{\lambda}) K\big(h^{-1}(\boldsymbol{X}_t - \boldsymbol{x})\big) = \mathbf{0}, \tag{3.15}$$

where $\mathbf{0}$ denotes the d-dimensional vector of zeros. At location $\boldsymbol{x}$ the probability mass is therefore not necessarily uniform, i.e., $(T - i_d - 1)^{-1}$ similar to (3.13), but shifted such that (3.15) holds. In addition,

$$p_t(\boldsymbol{x}; \boldsymbol{\lambda}) \geq 0 \quad \text{and} \quad \sum_{t=i_d+1}^{T-1} p_t(\boldsymbol{x}; \boldsymbol{\lambda}) = 1, \tag{3.16}$$

are imposed to guarantee that $p_t(\boldsymbol{x};\boldsymbol{\lambda})$ are indeed probabilities. These are needed to preserve the property of being a proper estimator of the conditional CDF. Hall and Presnell (1999) and Hall et al. (1999) propose to select the probabilities via the empirical likelihood by maximizing $\sum_{t=i_d+1}^{T-1}\ln(p_t(\boldsymbol{x};\boldsymbol{\lambda}))$, subject to the above constraints (3.15) and (3.16). Note that the strict positivity constraint of the probabilities is implicitly imposed by the objective function. Let $\boldsymbol{\lambda}=\boldsymbol{\lambda}(\boldsymbol{x})$ denote the d-dimensional vector of Lagrange-parameters λ_l, with $l=1,...,d$, for condition (3.15) of the reduced optimization problem. Then, after some algebra (see Li and Racine (2006, pp. 186–189)) the probabilities are given by

$$p_t(\boldsymbol{x};\boldsymbol{\lambda})=\frac{1}{(T-i_d-1)(1+h^{-d}\boldsymbol{\lambda}(\boldsymbol{X}_t-\boldsymbol{x})'K(h^{-1}(\boldsymbol{X}_t-\boldsymbol{x})))}, \tag{3.17}$$

with $\boldsymbol{\lambda}$ not having a closed form solution. Given (3.15) and (3.17) the optimization problem to determine the probabilities can be simplified leading to the maximization of $L(\boldsymbol{x};\boldsymbol{\lambda})=-\sum_{t=i_d+1}^{T-1}\ln(1+h^{-d}\boldsymbol{\lambda}(\boldsymbol{X}_t-\boldsymbol{x})'K(h^{-1}(\boldsymbol{X}_t-\boldsymbol{x})))$ with respect to $\boldsymbol{\lambda}$. The first order conditions read

$$\sum_{t=i_d+1}^{T-1}\frac{(\boldsymbol{X}_t-\boldsymbol{x})K(h^{-1}(\boldsymbol{X}_t-\boldsymbol{x}))}{1+h^{-d}\boldsymbol{\lambda}(\boldsymbol{X}_t-\boldsymbol{x})'K(h^{-1}(\boldsymbol{X}_t-\boldsymbol{x}))}=\mathbf{0}. \tag{3.18}$$

Given the optimal probabilities at $\boldsymbol{x}$, the weighted Nadaraya-Watson estimator is derived by minimizing (3.4), for $q=0$, resulting in

$$\widehat{m}_{wnw}(\boldsymbol{x})=\frac{\sum_{t=i_d+1}^{T-1}p_t(\boldsymbol{x};\boldsymbol{\lambda})K(h^{-1}(\boldsymbol{X}_t-\boldsymbol{x}))\phi(X_{t+1})}{\sum_{t=i_d+1}^{T-1}p_t(\boldsymbol{x};\boldsymbol{\lambda})K(h^{-1}(\boldsymbol{X}_t-\boldsymbol{x}))}, \tag{3.19}$$

where $\boldsymbol{\lambda}$ solves (3.18). In summary, $p_t(\boldsymbol{x};\boldsymbol{\lambda})>0$ and $K(h^{-1}(\boldsymbol{X}_t-\boldsymbol{x}))\geq 0$ the estimator in (3.19) preserves the property of being a proper estimator of the conditional CDF. Because of the condition in (3.15) the weighted Nadaraya-Watson estimator reproduces the superior bias properties of the local linear estimator.

The success of the weighted Nadaraya-Watson estimator is mainly owed to the fact that now a simple and proper estimator of the conditional CDF with considerable lower bias is available. Applications of the estimator are therefore focussed in the realm of estimating this distribution. Cai (2002) proposes a quantile regression estimator based on the weighted Nadaraya-Watson estimator. Kato (2012) constructs an estimator of the conditional expected shortfall. Cai and Wang (2008) consider a similar problem and the estimation of the conditional Value-at-Risk using

the weighted Nadaraya-Watson estimator. Bao et al. (2006) evaluate the predictive performance of the estimator in various Value-at-Risk models. Tay and Ting (2008) investigate the CDF of high-frequency price changes conditional on trading volume and duration between trades.

Given the success of the estimator, procedures to select the bandwidth are surprisingly scarce. For example, Bao et al. (2006) use standard cross-validation to determine the bandwidth. This procedure is not valid in case of time series data (see Arlot and Celisse (2010, pp. 65–66)). Kato (2012) and Tay and Ting (2008) use a legitimate bootstrap estimator of the mean squared error based on the autoregressive bootstrap first introduced in Hall et al. (1999). Cai (2002) uses for his examples a nonparametric version of Akaike's information criterion introduced in Cai and Tiwari (2000). This method will serve as a benchmark method to select the bandwidth and is introduced in the following section.

3.3 The benchmark method to select the bandwidth

Cai and Tiwari (2000) introduce a nonparametric version of the Akaike information criterion to select the bandwidth for the local linear estimator. Because the weighted Nadaraya-Watson estimator belongs to the class of linear smoothers, i.e., the smoothed values can be expressed as a linear transformation of the observed values, this bandwidth selection method remains valid. Cai (2002, p. 178) uses this method to select the bandwidth the weighted Nadaraya-Watson estimator to estimate regression quantiles. To determine the smoother matrix, $\mathbf{\Omega}$, let

$$\omega_t(\boldsymbol{x}) = \frac{p_t(\boldsymbol{x};\boldsymbol{\lambda})K(h^{-1}(\boldsymbol{X}_t - \boldsymbol{x}))}{\sum_{t=i_d+1}^{T-1} p_t(\boldsymbol{x};\boldsymbol{\lambda})K(h^{-1}(\boldsymbol{X}_t - \boldsymbol{x}))},$$

then the estimates for $\{\phi(X_{i_d+2}), ..., \phi(X_T)\}$ are given by

$$\begin{pmatrix} \widehat{\phi(X_{i_d+2})} \\ \widehat{\phi(X_{i_d+3})} \\ \vdots \\ \widehat{\phi(X_T)} \end{pmatrix} = \underbrace{\begin{pmatrix} \omega_{i_d+1}(\boldsymbol{X}_{i_d+1}) & \omega_{i_d+2}(\boldsymbol{X}_{i_d+1}) & ... & \omega_{T-1}(\boldsymbol{X}_{i_d+1}) \\ \omega_{i_d+1}(\boldsymbol{X}_{i_d+2}) & \omega_{i_d+2}(\boldsymbol{X}_{i_d+2}) & ... & \omega_{T-1}(\boldsymbol{X}_{i_d+2}) \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{i_d+1}(\boldsymbol{X}_{T-1}) & \omega_{i_d+2}(\boldsymbol{X}_{T-1}) & ... & \omega_{T-1}(\boldsymbol{X}_{T-1}) \end{pmatrix}}_{=:\mathbf{\Omega}} \begin{pmatrix} \phi(X_{i_d+2}) \\ \phi(X_{i_d+3}) \\ \vdots \\ \phi(X_T) \end{pmatrix}.$$

The nonparametric AIC to select the bandwidth for the weighted Nadaraya-Watson

estimator is defined as

$$AIC_c(h) = \log(rss) + \frac{(T - i_d - 1) + \mathrm{Tr}(\mathbf{\Omega\Omega}')}{(T - i_d - 1) - \mathrm{Tr}(\mathbf{\Omega\Omega}') - 2}, \tag{3.20}$$

with

$$rss = \sum_{t=i_d+1}^{T-1} (\phi(X_{t+1}) - \widehat{\phi(X_{t+1})})^2,$$

where Tr denotes the trace operator of a matrix. To determine the optimal bandwidth for this procedure, denoted by h_{aic}, $AIC_c(h)$ is numerically minimized over a suitable grid of bandwidth values. The criterion given in (3.20) is a nonparametric extension of the usual Akaike information criterion to determine, e.g., the order of a linear time series model (see, e.g., Brockwell and Davis (2003, p. 173)). The penalization here is defined as $\mathrm{Tr}(\mathbf{\Omega\Omega}')$, i.e., the sum of eigenvalues of $\mathbf{\Omega\Omega}'$. Depending on h it therefore provides an indication of the amount of fitting the smoothing matrix does. The degrees of freedom in (3.20) are equal to $\mathrm{Tr}(\mathbf{\Omega\Omega}')$ instead of $\mathrm{Tr}(\mathbf{\Omega})$, as suggested in Cai and Tiwari (2000), because if the estimates are not considered for the entire range of the data, $\mathbf{\Omega}$ is not a square matrix. This alternative choice of the degrees of freedom is legitimate (see Hastie and Tibshirani (1990, p. 52–55)) and avoids selecting a bandwidth driven by outlying observations.

3.4 Bandwidth selection via the local bootstrap procedure

This section provides a bootstrap estimator of the integrated mean squared error (IMSE) based on the local bootstrap procedure to select the bandwidth for the weighted Nadaraya-Watson estimator.

The local bootstrap, introduced by Paparoditis and Politis (2000), is used by the authors to estimate the sampling distribution of kernel estimators, in particular the distribution of the ordinary Nadaraya-Watson estimator. Their procedure is extended in the present manuscript to the weighted Nadaraya-Watson estimator and is defined in the following way. Consider the estimation of the distribution of

X_{t+1} conditional on $\boldsymbol{X}_t = \boldsymbol{x}$ using the weighted Nadaraya-Watson estimator, i.e.,

$$\widehat{F}_{g,X_{t+1}|\boldsymbol{X}_t}(\cdot|\boldsymbol{x}) = \frac{\sum_{s=i_d+1}^{T-1} p_s(\boldsymbol{x};\boldsymbol{\lambda}) K(g^{-1}(\boldsymbol{X}_s - \boldsymbol{x})) \mathbb{1}_{(-\infty,\cdot]}(X_{s+1})}{\sum_{s=i_d+1}^{T-1} p_s(\boldsymbol{x};\boldsymbol{\lambda}) K(g^{-1}(\boldsymbol{X}_s - \boldsymbol{x}))}, \tag{3.21}$$

with pilot bandwidth g. The estimator in (3.21) consistently estimates the predictive distribution $F_{X_{t+1}|\boldsymbol{X}_t}(\cdot|\boldsymbol{x})$ (see Cai (2001) and Steikert (2014)). Instead of replicating the entire time series the local bootstrap replicates $\{X^*_{t+1}, \boldsymbol{X}_t\}_{i_d+1}^{T-1}$ of the observed pairs $\{X_{t+1}, \boldsymbol{X}_t\}_{i_d+1}^{T-1}$ such that

$$X^*_{t+1} \sim \widehat{F}_{g,X_{t+1}|\boldsymbol{X}_t}(\cdot|\boldsymbol{X}_t), \tag{3.22}$$

where X^*_{t+1} is a random variable taking values in $\{X_{i_d+2}, ..., X_T\}$. The method avoids cumbersome estimation of the data generating process and is therefore considered a model free approach. It demands however the estimation of $T - i_d - 1$ conditional CDFs to generate the bootstrap pairs. For each $t = i_d + 1, ..., T - 1$, the distribution (3.21) varies and because of the whitening by windowing principle (see Hart (1996, pp. 117–119)) the bootstrap replicates $(X^*_{s+1}, \boldsymbol{X}_s)$ and $(X^*_{t+1}, \boldsymbol{X}_t)$ are asymptotically independent for $s \neq t$. Given this feature of the local bootstrap it suffices to employ this independent resampling scheme. The procedure works because the limiting distribution of $\widehat{m}_h(\boldsymbol{x})$ is Gaussian depending only on the marginal distribution of $\boldsymbol{X}_t$ and the conditional CDF of X_{t+1} given $\boldsymbol{X}_t$ (see Cai (2001, p. 311)).

To construct the bootstrap estimator of the mean squared error (MSE) first, note that in general the joint distribution, F, of $\{X_t\}_{t=1}^T$ is unknown and therefore the sampling distribution, $G_T(\cdot, F)$, of $\widehat{m}_h(\boldsymbol{x}) - m(\boldsymbol{x})$ is also unknown. Because the MSE of $\widehat{m}_h(\boldsymbol{x})$ is a function of $G_T(\cdot, F)$, namely $MSE(\widehat{m}_h(\boldsymbol{x})) = \int y^2 dG_T(y, F)$, it cannot be determined exactly. Hence, the distribution of $\widehat{m}_h(\boldsymbol{x}) - m(\boldsymbol{x})$ is approximated by means of the distribution of $\widehat{m}^*_h(\boldsymbol{x}) - m^*(\boldsymbol{x})$, where $\widehat{m}^*_h(\boldsymbol{x})$ and $m^*(\boldsymbol{x})$ are the bootstrap versions of $\widehat{m}_h(\boldsymbol{x})$ and $m(\boldsymbol{x})$, respectively. Given the above resampling mechanism $\widehat{m}^*_h(\boldsymbol{x})$ and $m^*(\boldsymbol{x})$ are determined in the following way. Because only the paris $\{X^*_{t+1}, \boldsymbol{X}_t\}_{i_d+1}^{T-1}$ are considered it immediately follows that

$$\widehat{m}^*_h(\boldsymbol{x}) = \frac{\sum_{t=i_d+1}^{T-1} p_t(\boldsymbol{x};\boldsymbol{\lambda}) K(h^{-1}(\boldsymbol{X}_t - \boldsymbol{x})) \phi(X^*_{t+1})}{\sum_{t=i_d+1}^{T-1} p_t(\boldsymbol{x};\boldsymbol{\lambda}) K(h^{-1}(\boldsymbol{X}_t - \boldsymbol{x}))}. \tag{3.23}$$

To determine the centering variable $\widehat{m}^*(\boldsymbol{x}) = \mathbb{E}^*(\phi(X^*_{t+1})|\boldsymbol{X}_t = \boldsymbol{x})$ note that, given

(3.21), the probability mass function for the random variable X^*_{t+1} given $\boldsymbol{X}_t$ reads

$$\mathbb{P}(X^*_{t+1} = X_{s+1}|\boldsymbol{X}_t) = \frac{p_s(\boldsymbol{X}_t)K(g^{-1}(\boldsymbol{X}_s - \boldsymbol{X}_t))}{\sum_{k=i_d+1}^{T-1} p_k(\boldsymbol{X}_t)K(g^{-1}(\boldsymbol{X}_k - \boldsymbol{X}_t))}, \tag{3.24}$$

for $s = i_d + 1, i_d + 2, ..., T - 1$ with

$$p_s(\boldsymbol{X}_t) = \frac{1}{(T - i_d - 1)(1 + g^{-d}\boldsymbol{\lambda}'(\boldsymbol{X}_s - \boldsymbol{X}_t)K(g^{-1}(\boldsymbol{X}_s - \boldsymbol{X}_t)))},$$

and unique

$$\boldsymbol{\lambda}' = \underset{\boldsymbol{\lambda}\in\mathbb{R}^d}{\arg\max}\left\{ - \sum_{k=i_d+1}^{T-1} \log\big(1 + g^{-d}\boldsymbol{\lambda}'(\boldsymbol{X}_k - \boldsymbol{X}_t)K\big(g^{-1}(\boldsymbol{X}_k - \boldsymbol{X}_t)\big)\big) \right\}.$$

Therefore,

$$\begin{aligned} m^*(\boldsymbol{x}) &= \mathbb{E}^*(\phi(X^*_{t+1})|\boldsymbol{X}_t = \boldsymbol{x}) \\ &= \sum_{s=i_d+1}^{T-1} \phi(X_{s+1})\mathbb{P}(X^*_{t+1} = X_{s+1}|\boldsymbol{X}_t = \boldsymbol{x}) \\ &= \sum_{s=i_d+1}^{T-1} \phi(X_{s+1}) \frac{p_s(\boldsymbol{x})K(g^{-1}(\boldsymbol{X}_s - \boldsymbol{x}))}{\sum_{k=i_d+1}^{T-1} p_k(\boldsymbol{x})K(g^{-1}(\boldsymbol{X}_k - \boldsymbol{x}))} \\ &= \widehat{m}_g(\boldsymbol{x}). \end{aligned} \tag{3.25}$$

In summary this implies that, to study the distribution of $\widehat{m}_h(\boldsymbol{x}) - m(\boldsymbol{x})$ the bootstrap approximation $\widehat{m}^*_h(\boldsymbol{x}) - \widehat{m}_g(\boldsymbol{x})$ is used and the bootstrap estimator of the MSE is defined as

$$\widehat{MSE}(\widehat{m}_h(\boldsymbol{x})) = \mathbb{E}\big((\widehat{m}^*_h(\boldsymbol{x}) - \widehat{m}_g(\boldsymbol{x}))^2 \big| X_1, ..., X_T\big), \tag{3.26}$$

with $\widehat{m}^*_h(\boldsymbol{x})$ and $\widehat{m}_g(\boldsymbol{x})$ given in (3.23) and (3.25), respectively.

3.4.1 Construction of the bootstrap sample

To replicate the pairs $\{X_{t+1}, \boldsymbol{X}_t\}_{t=i_d+1}^{T-1}$ the conditional CDF, $F_{X_{t+1}|\boldsymbol{X}_t}(\cdot|\boldsymbol{X}_t)$, is estimated using the weighted Nadaraya-Watson estimator given in (3.21). For a given pilot bandwidth g, each bootstrap value X^*_{t+1} is selected according to $\widehat{F}_{g,X_{t+1}|\boldsymbol{X}_t}(y|\boldsymbol{X}_t)$, with $y \in \{X_j\}_{j=i_d+2}^{T}$. The pilot bandwidth g determines the likelihood of each X_{t+1}

being selected, depending on the closeness of $\boldsymbol{X}_{i_d+1}, \boldsymbol{X}_{i_d+2}, ..., \boldsymbol{X}_{T-1}$ to $\boldsymbol{X}_t$ as seen in (3.24). To illustrate the selection mechanism consider the polynomial autoregressive (PAR) process

$$X_t = 3.76X_{t-1} - 0.235X_{t-1}^2 + 0.3\nu_t, \tag{3.27}$$

where the errors ν_t are independent for $s \neq t$, independent of $\{X_s\}_{s \leq t}$, and are uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$. The process is discussed in more detail in Section 3.5. Let $d = 1$, $i_1 = 2$ (3-step ahead prediction), $X_0 = 1$, $T = 2\,000$ and consider the deterministic version of the model in (3.27), i.e., $\nu_t = 0$ for all t. The gray curve in Figure I depicts the simulated sample. Furthermore, let $g = 1$ and $K(\cdot)$ be the standardized Epanechnikov kernel function to determine $\widehat{F}_{1,X_{t+1}|X_{t-2}}(y|X_{t-2})$, with $y \in \{X_j\}_{j=4}^T$, for the locations $X_{t-2} = 6$ and $X_{t-2} = 13$. Both estimated distributions are shown in Figure I. Note that $K(X_s - 6) = 0$ for all X_s for which $|X_s - 6| > 1$ (similar holds at $X_{t-2} = 13$). The support of $\widehat{F}_{1,X_{t+1}|X_{t-2}}(y|X_{t-2} = 6)$ therefore consists of those X_{s+1} for which the pairs $\{X_{s+1}, X_{s-2}\}_{s=3}^{T-1}$ exist given $X_{s-2} \in (5, 7)$ ($X_{s-2} \in (12, 14)$ for the other location). At $X_{t-2} = 13$ the support of the conditional CDF is larger than the support of $\widehat{F}_{1,X_{t+1}|X_{t-2}}(\cdot|X_{t-2} = 6)$ because more pairs $\{X_{s+1}, X_{s-2}\}_{s=3}^{T-1}$ exist for $X_{s-2} \in (12, 14)$ due to the form of the regression function.

[Insert Figure I about here]

To illustrate the influence of the pilot bandwidth g on the bootstrap sample consider Figure II. The gray lag-3 scatterplot represents a typical sample of the model given in (3.27). The left panel shows a typical bootstrap sample (black scatterplot) for a smaller value of the pilot bandwidth than for the bootstrap sample shown in the right panel. The bootstrap sample for the smaller value of g is closer to the original sample (gray scatterplot). A larger pilot bandwidth generates bootstrap samples with less pronounced structural features embedded in the original data.

[Insert Figure II about here]

The following subsection describes how the bandwidth is selected using the bootstrap estimator of the MSE given in (3.26). In addition, an iterative procedure is provided for which the selection of bandwidth is less dependent on the initial choice

of the pilot bandwidth g.

3.4.2 Choice of the bandwidth and the iterative procedure

Because the MSE is a pointwise measure of accuracy the bandwidth is selected via the bootstrap estimator of the integrated MSE, i.e., for given pilot bandwidth g,

$$h_{local} = \arg\min_{h>0} \int \mathbb{E}\big((\widehat{m}_h^*(\boldsymbol{x}) - \widehat{m}_g(\boldsymbol{x}))^2 \big| X_1, ..., X_T\big) d\boldsymbol{x}. \tag{3.28}$$

Thus, h_{local} is the bandwidth h that minimizes the expected squared difference between two estimated curves. To provide some insights in how the selection mechanism of h_{local} works let h_{emp} denote the (unknown) empirically optimal bandwidth minimizing the integrated squared difference between $\widehat{m}_h(\boldsymbol{x})$ and $m(\boldsymbol{x})$ for a given set of locations (for a formal definition see (3.35)). Suppose the pilot bandwidth g is reasonably close to h_{emp} and $g > h_{emp}$. Then, the estimated curve $\widehat{m}_g(\boldsymbol{x})$ over-smoothes the original data. In addition, the support of the estimated conditional CDFs to select the bootstrap values is large, eroding some of the structural features in the original data. Thus, theses features are less pronounced in the bootstrap sample than in the original sample (compare also Figure II). This implies, that for all $h \geq g$ the bootstrap estimate $\widehat{m}_h^*(\boldsymbol{x})$ is even more over-smoothed than $\widehat{m}_g(\boldsymbol{x})$ for the original data resulting in a large integrated MSE. To compensate the over-smoothing, h_{local} is a value strictly less than the pilot bandwidth g. This mechanism provides reason for an iterative procedure such that h_{local} depends less on the initial choice of g. The iteration process is similar to the one proposed in Faraway and Jhun (1990) where it is used to select the bandwidth for kernel density estimators. Let i denote the iteration step. For $i = 1$ select some (possibly large) initial pilot bandwidth g_1. Then, solving (3.28) results in bandwidth h_1 strictly less than g_1. For $i > 1$ set $g_i = h_{i-1}$ as long as $g_i > h_i$. If $g_i \leq h_i$ stop the iteration process and set $h_{local} = h_{i-1}$. The algorithm below provides all necessary steps to implement the iterative bandwidth selection method using the local bootstrap procedure.

Algorithm 3.4.1 (Iterated local bootstrap procedure). Let $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ denote the set of locations $\boldsymbol{x}$, then the iterated local bootstrap procedure consists of the following steps:

1. For given pilot bandwidth $g_i > 0$, compute the nonparametric estimate $\widehat{m}_{g_i}(\boldsymbol{x})$ using the observed sample $\{X_t\}_{t=1}^T$.

2. Given g_i, compute the conditional probability $\mathbb{P}(X^*_{t+1} = X_{s+1}|\boldsymbol{X}_t)$ for all $s, t = i_d + 1, i_d + 2, ..., T - 1$ according to (3.24).

3. Generate B bootstrap pairs $\{X^*_{t+1}, \boldsymbol{X}_t\}_{t=i_d+1}^{T-1}$ by bootstrapping X_{t+1} according to the probabilities of the previous step.

4. For each bootstrap sample $b = 1, ..., B$ of the previous step and given bandwidth h compute $\widehat{m}^{*(b)}_{h_i}(\boldsymbol{x})$ according to (3.23).

5. Select the bandwidth h_i such that

$$h_i = \underset{h>0}{\arg\min} \sum_{x_{i_1} \in \mathcal{X}_1} \cdots \sum_{x_{i_d} \in \mathcal{X}_d} \frac{1}{B} \sum_{b=1}^{B} \Big(\widehat{m}^{*(b)}_h(x_{i_1}, ..., x_{i_d}) - \widehat{m}_{g_i}(x_{i_1}, ..., x_{i_d}) \Big)^2. \tag{3.29}$$

6. For further iterations set $g_{i+1} = h_i$ and repeat steps 1 to 5 until $g_i \leq h_i$. Choose $h_{local} = h_{i-1}$.

Choosing the initial pilot bandwidth g_1 reasonably well, i.e., close but greater than h_{emp} reduces the number of iterations and therefore computing time. A first possible suggestion, for $d = 1$, is to the set g_1 equal to a multiple of the maximum absolute difference of the sorted observed time series. This value guarantees the existence of $\widehat{m}_g(\boldsymbol{x})$ and is usually larger than h_{emp}. A second, computationally more intense, possibility is to compute h_{aic} according to the method discussing in Section 3.3 and set $g_1 = 2h_{aic}$ or any other suitable multiple of h_{aic}. The results of the simulation study below are robust regarding the selection of g_1. Thus, the conclusions do not alter for either choice of the initial pilot bandwidth.

A (theoretically) more sophisticated value for g_1 is to consider the minimization of the integrated MSE of the second derivative of $\widehat{m}(\boldsymbol{x})$, i.e., $\int \mathbb{E}(\widehat{m}''_g(\boldsymbol{x}) - m''(\boldsymbol{x}))^2 d\boldsymbol{x}$. This approach achieves the right rate of decay for g (see Hall et al. (1995, p. 1927) for a similar approach). The problem, however, is that derivative estimators for the weighted Nadaraya-Watson estimator have not been developed so far.

Concerning the number of bootstrap samples B the examples below work well for $B = 1\,000$ or even less. Results not shown indicate that choices of B as large as 100 do not alter the decision to select the bandwidth. In any case, making use of the fact that the procedure bootstraps the pairs $\{X_{t+1}, \boldsymbol{X}_t\}_{t=i_d+1}^{T-1}$ and not the entire series $\{X_t\}_{t=1}^{T}$ the procedure in capable of coping with a large number of bootstrap samples. Making use of fact that the estimator is evaluated for the

bootstrap pairs $\{X^*_{t+1}, \boldsymbol{X}_t\}_{t=i_d+1}^{T-1}$ drastically reduces computing time because the probabilities $p_t(\boldsymbol{x}; \boldsymbol{\lambda})$ as well as kernel weights have to be determined only once and not for each sample of bootstrap pairs separately (see also (3.23)). This implies also that $\boldsymbol{\lambda}$ is determined only once by numerical methods. Evaluating $m_h^{*(b)}(x_{i_1}, ..., x_{i_d})$ for $b = 1, ..., B$ with $B = 100$ is therefore not much different than for $B = 1\,000$. The only difference is the summation $\sum_{t=i_d+1}^{T-1} p_t(\boldsymbol{x}; \boldsymbol{\lambda}) K(h^{-1}(\boldsymbol{X}_t - \boldsymbol{x}))\phi(X^*_{t+1})$ which is promptly executed.

3.5 Numerical examples

To examine the performance of the bandwidth selection method for various statistics I consider common examples of univariate nonlinear time series. As a benchmark, I provide results based on the nonparametric version of Akaike's information.

3.5.1 Simulation framework

Härdle (1992, p. 247) names the threshold AR, exponential AR, smooth transition AR, random coefficient, bilinear, and the autoregressive conditional heteroscedastic models as the most common nonlinear time series models. The simulation study therefore shall consist of most of these models while adding two further models. Given Härdle's list, the threshold AR (see Tong and Lim (1980)) which is in general not a smooth model is omitted from the study because the estimator requires smoothness (see Section 3.2). Also the random coefficient model, which in its purest form is a white noise model (see, e.g., Hamilton (1994, pp. 372–377) as well as Paparoditis and Politis (2000, p. 147) for an example), is omitted. For this type of model it is conjectured that the proposed selection method will not produce satisfactory results because a clear signal is missing. The bilinear model (see Subba Rao (1981)) is also omitted because it is difficult to develop a comprehensive estimation theory for it, which seems to be a reason why this type of model is not often considered in the literature. To extent the above list the polynomial AR model and the rational nonlinear AR model are added. In particular, the following data generating

processes are considered:

$$\text{(EAR)} \quad X_t = \left(0.26 + 7.99e^{-3X_{t-1}^2}\right) X_{t-1} + 0.5\varepsilon_t \tag{3.30}$$

$$\text{(PAR)} \quad X_t = 3.76X_{t-1} - 0.235X_{t-1}^2 + 0.3\nu_t \tag{3.31}$$

$$\text{(RNLAR)} \quad X_t = \frac{25(X_{t-1} - 1)}{1 + X_{t-1}^2} + 2.4\varepsilon_t \tag{3.32}$$

$$\begin{aligned}\text{(STAR)} \quad X_t &= 0.1(X_{t-1} - 0.1) \\ &\quad + 0.9(X_{t-1} - 0.1)\tanh(X_{t-1} - 0.1) + 0.7\varepsilon_t\end{aligned} \tag{3.33}$$

$$\begin{aligned}\text{(NLAR-TVV)} \quad X_t &= \frac{1}{1 + e^{-X_{t-1}}} + \sigma_t\nu_t, \\ &\text{with } \sigma_t^2 = \psi(X_{t-1} - 1.2) + \frac{3}{2}\psi(X_{t-1} + 1.2),\end{aligned} \tag{3.34}$$

with ε_t and ν_t being independent standard Gaussian respectively independent uniform distributed on $[-\sqrt{3}, \sqrt{3}]$ and both are independent of $\{X_s\}_{s \le t}$. Furthermore, $\psi(\cdot)$ denotes the standard Gaussian probability density function.

The exponential autoregressive (EAR) model given in (3.30), first introduced to model nonlinear random vibrations (see Ozaki (1980, 1982) and Haggan and Ozaki (1981)) and later refined and applied to topics other than physics (see, e.g., Teräsvirta (1994)), is a special case of the so-called univariate exponential smooth transition AR model. The model is a responds to the lack of smoothness of traditional threshold models. The process moves similar to a simple AR(1) process with parameter 0.26 for large values of $|X_{t-1}|$. For small $|X_{t-1}|$, however, the AR coefficient is roughly $0.26 + 7.99$. Note that (3.30) is stationary because $0.26 + 7.99e^{-3x^2} \to 0.26$ as $|x| \to \infty$ and $0.26 < 1$ (Foster-Lyapunov criterion, for a different criterion see Chan and Tong (1994, p. 305)). A typical time series data plot of 100 observations is presented in the top left panel of Figure III below.

The polynomial autoregressive (PAR) model in (3.31) is an example taken from Hall et al. (1999, p. 157) and Fan and Yao (2005, p. 445). Given the bounded support of the distribution of ν_t, the model is not explosive for $X_0 = 1$ with probability one (see Chan and Tong (1994, Theorem 1)). Applications of the PAR model can be found, e.g., in Cox (1977) and Chan and Tong (1994). A typical time series data plot of the model is given in the top right panel of Figure III.

The model given in (3.32) is an examples of a rational nonlinear autoregressive (RNLAR) model. Other examples of this type of model can be found in Granger and Lee (1999, p. 263), Fan and Gijbels (1996, p. 221), and Fan and Yao (2005, p. 16). Here a different model is selected because the first

model in the aforementioned reference is not smooth and for the second and third model the distribution of the noise exhibits bounded support which represents an additional constraint in their models. Thus, (3.32) is more demanding due to the larger support of the distribution of the noise component. Stationarity is guaranteed as long as the polynomial in the denominator is of larger order than in the numerator (Foster-Lyapunov criterion). The mid left panel of Figure III shows a typical plot of the time series data.

To represent the class of smooth transition autoregressive (STAR) models a model suggested by Bacon and Watts (1971, p. 528) with a hyperbolic tangent function as a transition function between two linear regimes is considered in (3.33). An alternative transition function is the logistic transition function for which $(1 + \exp(-X_{t-1}))^{-1}$ is an example of. The Foster-Lyapunov criterion may serves again for proving stationarity for (3.33). A plot of the time series data is given in the mid right panel of Figure III.

The last model, the nonlinear autoregressive time varying volatility (NLAR-TVV) model, is an example taken from Härdle (1992, p. 253). Note the similarities to the ARCH(1) model once the term $(1 + \exp(-X_{t-1}))^{-1}$ is omitted. For a simple ARCH(1) nonparametric estimation techniques are not necessary due to the linear structure of the statistic. Thus, a model with a nonlinear structure is chosen. The bottom left panel of Figure III shows a typical time series data plot of the model.

[Figure III about here]

For all of the above models, except the PAR and the NLAR-TVV model, I consider the estimation of the one-step ahead prediction, i.e., I estimate $m(x) = \mathbb{E}(X_{t+1}|X_t = x)$ using (3.19). For the PAR and NLAR-TVV model I consider the estimation of the three-step ahead prediction and the one-step ahead prediction of the conditional variance, respectively. For the latter model this results is estimating $\mathrm{Var}(X_{t+1}|X_t = x) = \mathbb{E}(X_{t+1}^2|X_t = x) - (\mathbb{E}(X_{t+1}|X_t = x))^2$ via $\widehat{m}_h(x^2) - (\widehat{m}_h(x))^2$ using one bandwidth. To gain insides about the statistics that are estimated Figure IV presents scatterplots of typical samples of the models given in (3.30)–(3.33) as well as the regression being estimated (solid lines). The bottom left panel shows the true predictive variance function of (3.34). The figure highlights the nonlinear structures of the functions being estimated. All functions are estimated on the set $\mathcal{X}$ which does not coincide with the entire range of the data. Otherwise the bandwidth is mostly driven by outlying observations in the sense that the first

bandwidth for which the estimator exists is optimal.

[Figure IV about here]

For each example I follow the same procedure. A sample of size T of the data generating process is simulated and the *empirically optimal* bandwidth, h_{emp}, of the estimate $\widehat{m}_h(x)$ of $m(x)$, given the set of locations $\mathcal{X}$, using the weighted Nadaraya-Watson estimator is computed via

$$h_{emp} = \arg\min_{h>0} \sum_{x \in \mathcal{X}} (\widehat{m}_h(x) - m(x))^2. \tag{3.35}$$

The exact parameters used for each example, such as sample size and the set of locations are listed in Table I in the appendix. Regarding the sample size I consider small ($T = 400$) as well as large sizes ($T = 2\,000$). For the estimation I employ the standardized Epanechnikov kernel function given by $K(u) = 0.75(1 - u^2)\mathbb{1}_{\{|u|<1\}}$.

Algorithm 3.4.1 determines h_{local} using a Monte-Carlo simulation with $B = 1\,000$ to approximate the conditional expectation in (3.28). Furthermore, I determine h_{aic} according to the discussion in Section 3.3 as a benchmark.

I measure the performance of each selection method by the mean absolute deviation error (MADE), i.e., for $h \in \{h_{emp}, h_{local}, h_{aic}\}$. It is defined as the mean absolute difference between the estimated and true regression function for a given set of locations $\mathcal{X}$, i.e.,

$$\text{MADE}(h) = \frac{1}{\#\mathcal{X}} \sum_{x \in \mathcal{X}} \left|\widehat{m}_h(x) - m(x)\right|, \tag{3.36}$$

with $h \in \{h_{emp}, h_{aic}, h_{local}\}$. The MADE is a common criterion to measure the performance of bandwidth selection methods (see, e.g., Hall et al. (1999) and Fan and Gijbels (1996, pp. 80–83)). It weights each error evenly and is therefore not driven by large errors such as the (integrated) MSE.

For the ease of reading, I abbreviate the MADE of the estimated curve, using the weighted Nadaraya-Watson with a particular bandwidth h, by its bandwidth. Thus, the MADE of h_{emp} refers to the MADE of the curve estimate $\widehat{m}_h(x)$ for all $x \in \mathcal{X}$ using $h = h_{emp}$. Note that I also compute the MADE for the empirically optimal bandwidth h_{emp} since it represents the best nonparametric fit of the data, given the square loss function in (3.35). This allows to compare the results of the

other bandwidth selection methods to the (unknown) best fit possible.

The above steps are repeated a separate Monte Carlo simulation with 1 000 trials to determine the distribution of outcomes of MADEs. Further, I provide kernel density estimates of the MADE deviation, i.e., for each simulated sample I compute MADE(h) – MADE(h_{emp}) for $h \in \{h_{aic}, h_{local}\}$.

3.5.2 Results

In terms of the mean absolute deviation error, defined in (3.36), the box plots of Figure V show that the PAR and RNLAR model produce a relatively large MADE for the empirically optimal bandwidth h_{emp}. For the PAR model this is owed to the dispersed data for values around $X_{t-2} = 8$ (see top right panel of Figure IV). The dispersion itself causes a larger error but also implies a larger bandwidth to guarantee the existence of the estimator. This in turn over-smoothes data at extremes of the regression function such as for values around $X_{t-2} = 5$ and $X_{t-2} = 13.5$. This slight over-smoothing produces the larger MADE for h_{emp} due to the large amount of data at these points. A similar argument holds for the RNLAR model. Here the slight over-smoothing produces a large deviation at $X_t = 0$ at the same time existence is compromised for smaller bandwidth because of the dispersed data around $X_t = -20$.

The box plots of Figure V reveal that the local bootstrap procedure outperforms the benchmark in all of the considered examples. Furthermore, the box plots of h_{emp} and h_{local} are very close, implying similar distributions of the outcomes of MADEs. That is, the median marker, upper and lower fences, as well as the outliers for the box plots almost coincide. The nonparametric version of Akaike's information criterion is for all simulation, except for the RNLAR model, not competitive to the local bootstrap. In particular, for the PAR and STAR model the selection of bandwidth produces a larger variability in the MADE outcomes. Furthermore, the benchmark method produces much more outliers than the local bootstrap procedure.

[Figure V about here]

The last figure, Figure VI, presents kernel density estimates of MADE(h) – MADE(h_{emp}), for $h \in \{h_{aic}, h_{local}\}$. This gives a one-to-one comparison of each bandwidth selection method on the basis of each simulated data set. I employ Silverman's rule-of-thumb to select the bandwidth for the kennel density estimates (see Li and Racine (2006, p. 14)). The figure emphasizes that in all cases the local bootstrap selects the bandwidth such that the MADE of h_{local} and h_{emp} are

close. All densities are slightly positively skewed but close to zero exhibiting very low variances. The results for the RNLAR model are remarkably sharp because most deviations are close to zero. The support of the densities for the benchmark method are much larger compared to the local bootstrap procedure. Also the densities are more positively skewed and further apart from zero. This implies a systematical larger error when using the nonparametric Akaike's information criterion. Since the densities feature more probability mass in the right tail larger errors occurred for some of the data sets. Note that the comparison based on each data set separately shows also that the benchmark method is inferior also for the RNLAR model. This conclusion was difficult to make for Figure V.

[Figure VI about here]

In summary, the simulation study shows that the proposed local bootstrap procedure selects the bandwidth for the weighted Nadaraya-Watson estimator such that it produces only a small differences between MADE(h_{local}) and MADE(h_{emp}).

3.6 Conclusion

The selection of bandwidth for nonparametric estimators such as local polynomial estimators and the weighted Nadaraya-Watson estimator is a crucial part of the estimation procedure. The bandwidth, as a free parameter, mainly determines the estimated model and deviations from the optimal bandwidth can lead to false conclusions regarding the estimate. There exist numerous bandwidth selection methods for the case of independent data; however, for time series data the methods are scarce. Existing methods based on parametric modeling rely on the full range of estimation techniques for nonlinear time series data. These methods are prone to select a non-optimal bandwidth without prior knowledge of the data generating process. A simpler method that is based on a nonparametric version of Akaike's information criterion is easy to implement but can produce unsatisfactory results.

In this manuscript a fully data-driven method to select the bandwidth for the weighted Nadaraya-Watson estimator is proposed. This easy-to-implement method is based on the local bootstrap procedure and is free of any parametric nonlinear estimation techniques. The method is therefore appealing for practical purposes, in particular for practitioners with little background in estimation of nonlinear time

series. The performance of the selection method is tested in a simulation study. Various common nonlinear time series models are investigated for which the conditional mean and conditional variance functions are estimated using the weighted Nadaraya-Watson estimator. The results of the study indicate that the proposed local bootstrap procedure to select the bandwidth for the weighted Nadaraya-Watson estimator is an appealing choice among the scarce list of bandwidth selection methods. The method selects the bandwidth such that the mean absolute deviation between the estimated curve and the unknown true curve is small. It outperforms the benchmark method, based on a nonparametric version of Akaike's information criterion, in all the considered examples.

Acknowledgements

I thank Silvia Grätz, Michael Wolf, and Jan Wrampelmeyer for valuable comments and suggestions.

Appendix

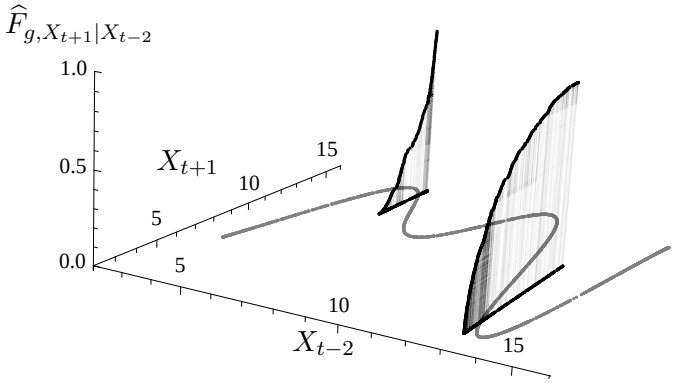
A Tables and Figures

Table I: Parameters used and statistics estimated in the simulation study for the models given in (3.30)–(3.34).

Model	Statistic	T	$\mathcal{X}$
EAR	$\mathbb{E}(X_{t+1}\|X_t = x)$	500	$[2.5\%, 97.5\%]$
PAR	$\mathbb{E}(X_{t+1}\|X_{t-2} = x)$	2 000	$[5\%, 95\%]$
RNLAR	$\mathbb{E}(X_{t+1}\|X_t = x)$	700	$[5\%, 95\%]$
STAR	$\mathbb{E}(X_{t+1}\|X_t = x)$	500	$[5\%, 95\%]$
NLAR-TVV	$\mathrm{Var}(X_{t+1}\|X_t = x)$	400	$[2.5\%, 97.5\%]$

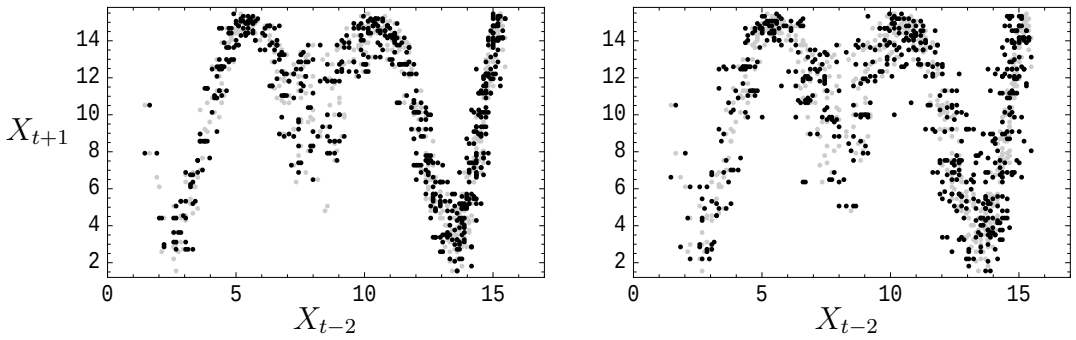
Notes: T and $\mathcal{X}$ denote the size of the simulated sample and the set of locations x. The range at which the estimator is evaluated is given by the quantile range of the data $\{X_t\}_{t=1}^{T}$. The statistic for the NLAR-TVV model is defined as $\mathrm{Var}(X_{t+1}|X_t = x) = \mathbb{E}(X_{t+1}^2|X_t = x) - (\mathbb{E}(X_{t+1}|X_t = x))^2$.

Figure I: An illustration of the selection of bootstrap values for the local bootstrap procedure.



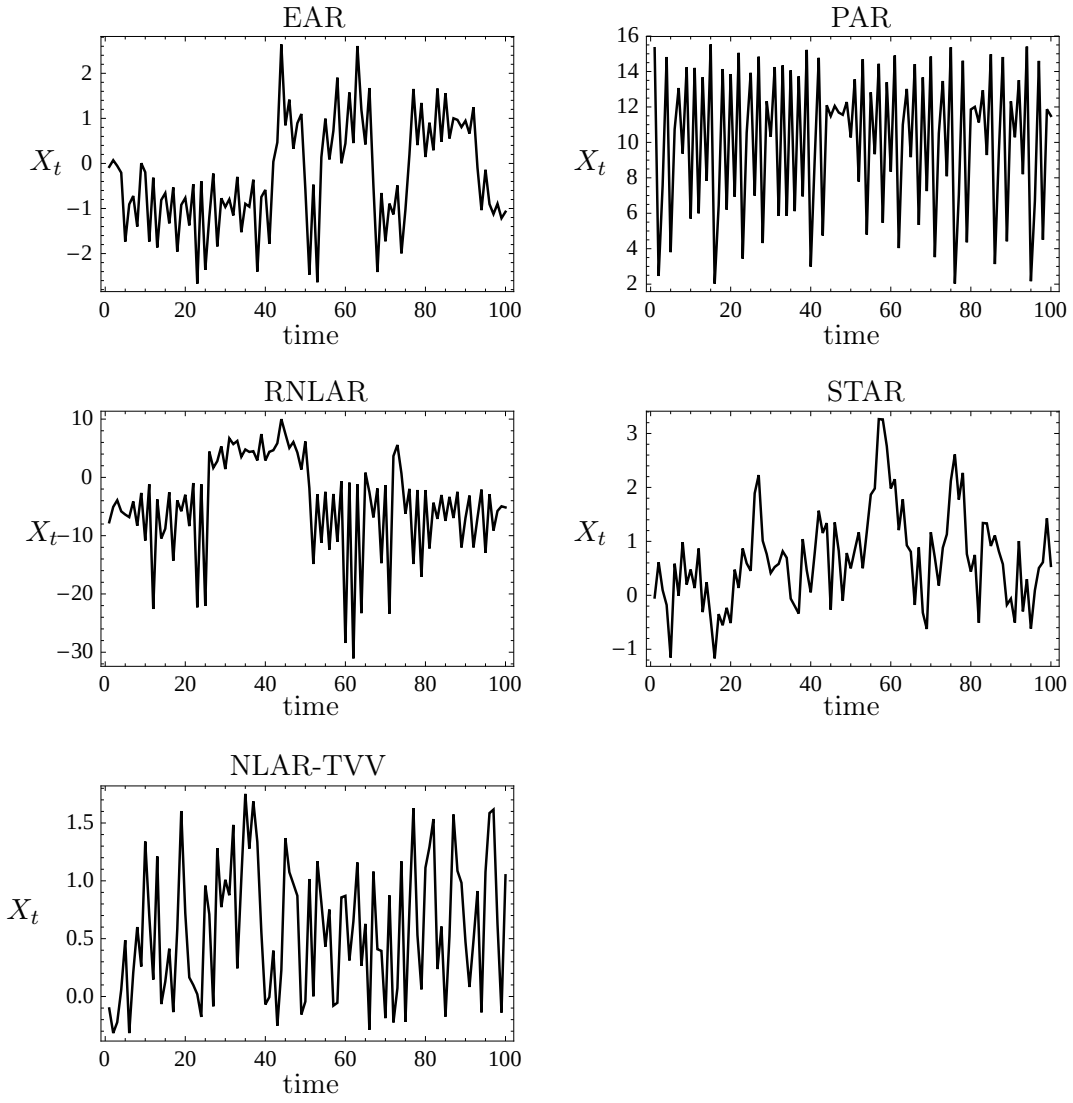
Notes: The grey curve represents a simulation of the 3-step ahead prediction of the deterministic version of the polynomial autoregressive model given in (3.27) with $X_0 = 1$ almost surely. The two estimated CDFs conditional on $X_{t-2} = 6$ and $X_{t-2} = 13$ determine the likelihood of X_{t+1} being selected for the bootstrap value X^*_{t+1}. The CDFs are estimated using the standardized Epanechnikov kernel with bandwidth $g = 1$.

Figure II: Comparison of bootstrap data for two different values of the pilot bandwidth g.



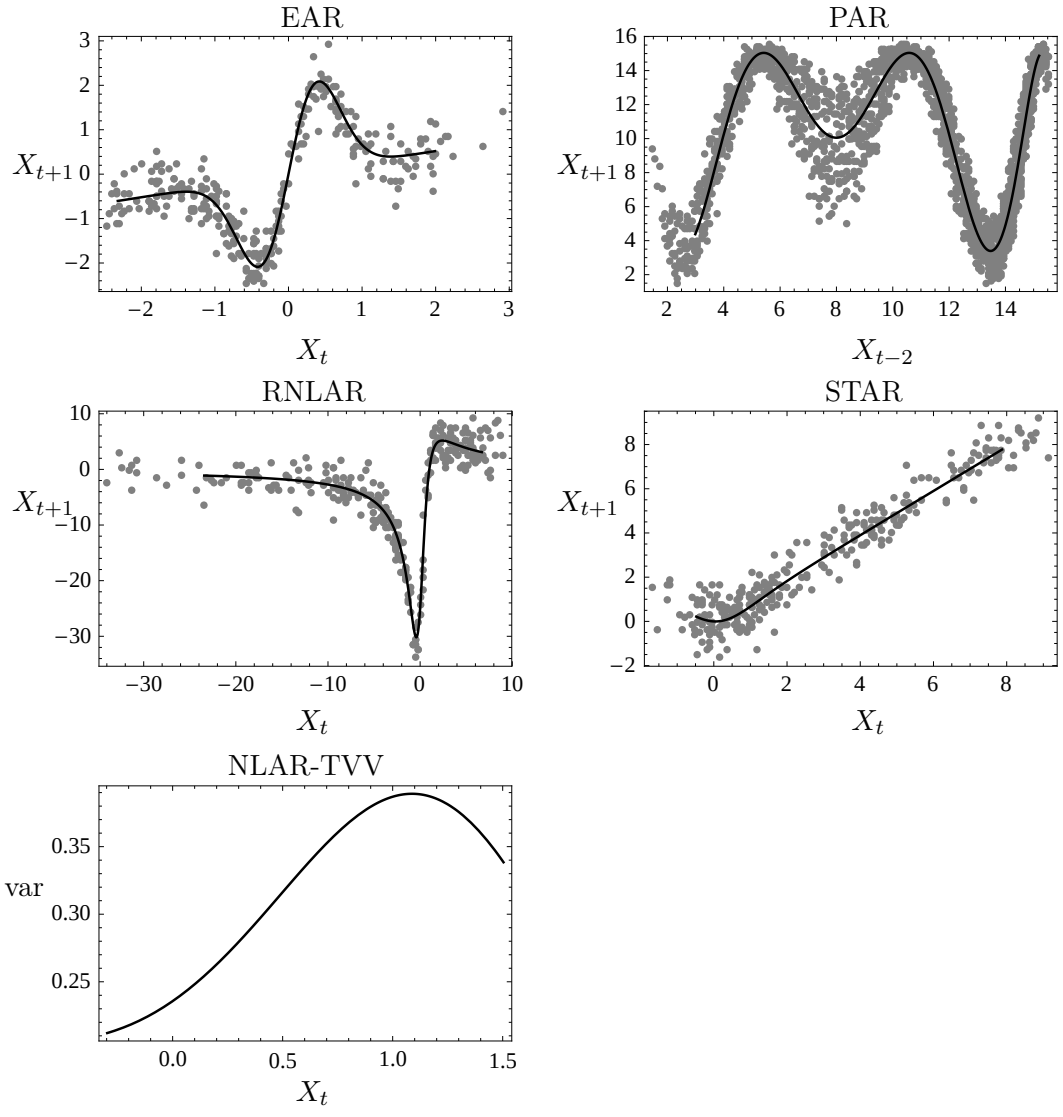
Notes: Both panels show the lag-3 scatterplot of a typical simulated data set of the model given in (3.27) (gray scatterplot) and an associated bootstrap sample (black scatterplot) for different values of the pilot bandwidth. Left panel: a typical bootstrap sample with pilot bandwidth $g = 0.4$. Right panel: a typical bootstrap sample with $g = 1$.

Figure III: Plots of typical time series data of the data generating processes given in (3.30)–(3.34).



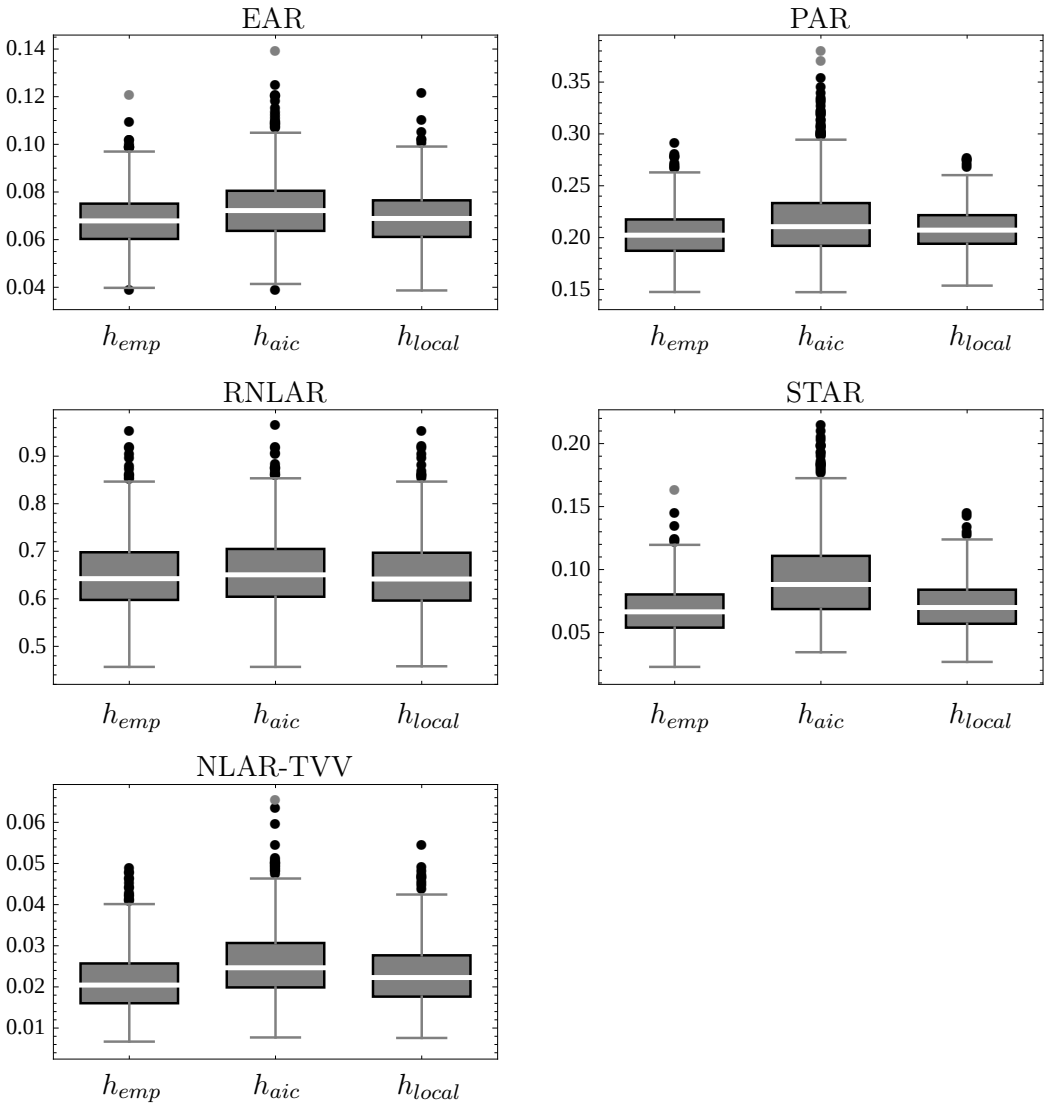
Notes: Top row: the exponential autoregressive (EAR) and polynomial autoregressive (PAR) model given in (3.30) and (3.31). Mid row: the rational nonlinear autoregressive (RNLAR) and smooth transition autoregressive (STAR) model given in (3.32) and (3.33). Bottom row: the nonlinear time varying autoregressive (NLAR-TVV) model.

Figure IV: Typical scatterplots of lagged data and the true regression functions of the models given in (3.30)–(3.34).



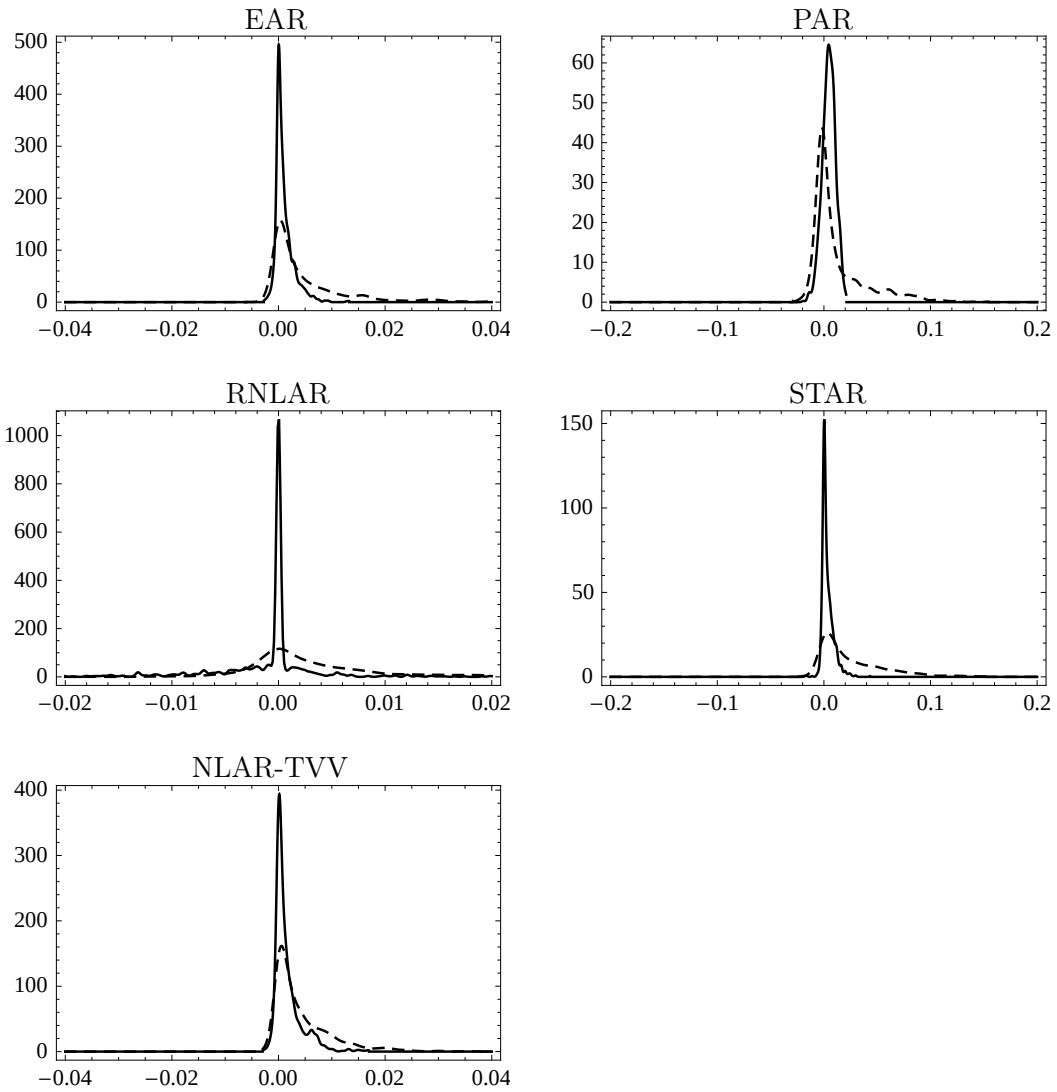
Notes: Top row: a typical scatterplots of lagged data and the true regression function (solid line) of the exponential autoregressive (EAR) and polynomial autoregressive (PAR) model, respectively. Mid row: rational nonlinear autoregressive (RNLAR) and smooth transition autoregressive (STAR) model. Bottom row: true conditional variance function of the nonlinear time varying autoregressive (NLAR-TVV) model, with var = $\mathrm{Var}(X_{t+1}|X_t)$.

Figure V: The resulting box plots of the mean absolute deviation errors (MADEs) of the simulation study.



Notes: Box plots of the mean absolute deviation errors (MADEs), defined in (3.36), for the empirically optimal bandwidth (h_{emp}, representing the infeasible benchmark), the nonparametric Akaike information criterion (h_{aic}, the benchmark), and the local bootstrap procedure (h_{local}). Top row: results for the exponential autoregressive (EAR) and polynomial autoregressive (PAR) model. Mid row: results for the rational nonlinear autoregressive (RNLAR) and smooth transition autoregressive (STAR) model. Bottom row: results for the nonlinear time varying autoregressive (NLAR-TVV) model.

Figure VI: The resulting kernel density estimates of the MADEs deviation of the simulation study.



Notes: Kernel density estimates of the MADE's deviation defined as MADE(h_{local}) – MADE(h_{emp}) (solid curve) and MADE(h_{aic}) – MADE(h_{emp}) (dashed curve). The bandwidth is selected according to Silverman's rule-of-thumb. Top row: results for the exponential autoregressive (EAR) and polynomial autoregressive (PAR) model. Mid row: the rational nonlinear autoregressive (RNLAR) and smooth transition autoregressive (STAR) model. Bottom row: results for the nonlinear time varying autoregressive (NLAR-TVV) model.

References

Arlot, S. and A. Celisse (2010): "A Survey of Cross-Validation Procedures for Model Selecton," *Statistics Surveys*, 4, 40–79.

Bacon, D. W. and D. G. Watts (1971): "Estimating the Transition between Two Intersecting Straight Lines," *Biometrika*, 58, 525..534.

Bao, Y., T.-H. Lee, and B. Saltoğlu (2006): "Evaluating Predictive Performance of Value-at-Risk Models in Emerging Markets: A Reality Check," *Journal of Forecasting*, 25, 101–128.

Brockwell, P. J. and R. A. Davis (2003): *Introduction to Time Series and Forecasting*, New York: Springer, 2 ed.

Cai, Z. (2001): "Weighted Nadaraya-Watson Regression Estimation," *Statistics & Probability Letters*, 51, 307–318.

——— (2002): "Regression Quantiles for Time Series," *Econometric Theory*, 18, 169–192.

Cai, Z. and R. C. Tiwari (2000): "Application of a Local Linear Autoregressive Model to BOD Time Series," *Environmetrics*, 11, 341–350.

Cai, Z. and X. Wang (2008): "Nonparametric Estimation of Conditional VaR and Expected Shortfall," *Journal of Econometrics*, 147, 120–130.

Chan, K. S. and H. Tong (1994): "A Note on Noisy Chaos," *Journal of the Royal Statistical Society. Series B*, 56, 301–311.

Chu, C.-K. and J. S. Marron (1991): "Choosing a Kernel Regression Estimator," *Statistical Science*, 6, 404–419.

Cleveland, W. S. (1979): "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Cox, D. R. (1977): "Discussion of Papers by Campbell and Walker, Tong and Morris," *Journal of the Royal Statistical Society: Series A*, 140, 453–454.

Fan, J. (1993): "Local Linear Regression Smoothers and their Minimax Efficiencies," *The Annals of Statistics*, 21, 196–216.

Fan, J., T. Gasser, I. Gijbels, M. Brockmann, and J. Engel (1995): "On Nonparametric Estimation via Local Polynomial Regression," *Working paper University of Louvain.*

Fan, J. and I. Gijbels (1992): "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 29, 2008–2036.

——— (1996): *Local Polynomial Modelling and its Applications*, Boca Raton: Chapman & Hall/CRC.

Fan, J. and Q. Yao (2005): *Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer.

Faraway, J. J. and M. Jhun (1990): "Bootstrap Choice of Bandwidth for Density Estimation," *Journal of the American Statistical Association*, 85, 1119–1122.

Granger, C. W. J. and T.-H. Lee (1999): "The Effect of Aggregation on Nonlinearity," *Econometric Review*, 18, 259–269.

Gu, J., Q. Li, and J.-C. Yang (2013): "Multivariate Local Polynomial Kernel Estimators: Leading Bias and Asymptotic Distribution," *Working paper Texas A&M University.*

Haggan, V. and T. Ozaki (1981): "Modelling Nonlinear Random Vibrations Using an Amplitude-Dependent Autoregressive Time Series Model," *Biometrika*, 68, 189–196.

Hall, P., S. N. Lahiri, and J. Polzehl (1995): "On Bandwidth Choice in Nonparametric Regression with both Short- und Long-Range Dependent Errors," *The Annals of Statistics*, 6, 1921–1936.

Hall, P. and B. Presnell (1999): "Intentionally Biased Bootstrap Methods," *Journal of the Royal Statistical Society: Series B*, 61, 143–158.

Hall, P., R. C. L. Wolff, and Q. Yao (1999): "Methods for Estimating a Conditional Distribution Function," *Journal of the American Statistical Association*, 94, 154–163.

Hamilton, J. D. (1994): *Time Series Analysis*, Princeton: Princeton University Press.

Härdle, W. (1992): *Applied nonparametric regression*, Cambridge: Cambridge University Press.

Hart, J. D. (1996): "Some Automated Methods of Smoothing Time-Dependent Data," *Nonparametric Statistics*, 6, 115–142.

Hastie, T. and C. Loader (1993): "Local Regression: Automatic Kernel Carpentry," *Statistical Science*, 8, 120–143.

Hastie, T. J. and R. J. Tibshirani (1990): *Generalized Additive Models*, London: Chapman & Hall/CRC.

Kato, K. (2012): "Weighted Nadaraya-Watson Estimation of Conditional Expected Shortfall," *Journal of Financial Econometrics*, 10, 265–291.

Lahiri, S. N. (2003): *Resampling Methods for Dependent Data*, New York: Springer.

Li, Q. and S. Racine (2006): *Nonparametric Econometrics. Theory and Practice*, Princeton: Princeton University Press.

Masry, E. (1996a): "Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates," *Journal of Time Series Analysis*, 17, 571–599.

——— (1996b): "Multivariate Regression Estimation: Local Polynomial Fitting for Time Series," *Stochastic Processes and their Application*, 65, 81–101.

Nadaraya, E. A. (1964): "On Estimating Regression," *Theory of Probability and Its Applications*, 9, 141–142.

Ozaki, T. (1980): "Non-Linear Time Series Models for Non-Linear Random Vibrations," *Journal of Applied Probability*, 17, 84–93.

——— (1982): "The Statistical Analysis of Perturbed Limit Cycle Processes using Nonlinear Time Series Models," *Journal of Time Series Analysis*, 3, 29–41.

Paparoditis, E. and D. N. Politis (2000): "The Local Bootstrap for Kernel Estimators under General Dependence Conditions," *Annals of the Institute of Statistical Mathematics*, 52, 139–159.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*, New York: Springer.

Ruppert, D. and M. P. Wand (1994): "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370.

Steikert, K. U. (2014): "The Weighted Nadaraya-Watson Estimator: Pointwise Strong Consistency and Convergence Rates for Strongly Mixing Processes," *Working paper University of Zurich.*

Stone, C. J. (1977): "Consistent Nonparametric Regression," *The Annals of Statistics*, 5, 595–645.

Subba Rao, T. (1981): "On the Theory of Bilinear Time Series Models," *Journal of the Royal Statistical Society. Series B*, 43, 244–255.

Tay, A. S. and C. Ting (2008): "Intraday Stock Prices, Volume, and Duration: A Nonparametric Conditional Density Analysis," in *High frequency financial econometrics: recent developments*, ed. by L. Bauwens, W. Pohlmeier, and D. Veredas, Heidelberg: Physica-Verlag, 253–268.

Teräsvirta, T. (1994): "Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models," *Journal of the American Statistical Association*, 89, 208–218.

Tong, H. and K. S. Lim (1980): "Threshold Autoregression, Limit Cycles and Cyclical Data," *Journal of the Royal Statistical Society. Series B*, 42, 245–292.

Watson, G. S. (1964): "Smooth Regression Analysis," *Sankhyā*, 26, 359–372.

Yu, K. and M. C. Jones (1998): "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228–237.

Curriculum Vitae

Kristoph U. Steikert

Birthday October 29, 1977

Academic education

2002–2005	Studies in economics at the University of Bonn
2005–2007	Master of Advanced Studies in Finance at the Eidgenössische Technische Hochschule (ETH) Zurich and the University of Zurich
2007–2014	Doctoral studies in banking and finance at the University of Zurich and the Swiss Finance Institute